

In Memoriam Rainald K. Bauer

*** 17.5 1931 † 2.10.1990**

Rainald K. Bauer wurde in München geboren, verbrachte im Münchener Raum seine Jugend und studierte an der Ludwig-Maximilians-Universität Mathematik. Nach seinem Studium blieb er noch zwei Jahre als Mitarbeiter der ökonometrischen Arbeitsgruppe am Ifo-Institut für Wirtschaftsforschung in München.

1956 begann seine Exilantenzeit außerhalb der bayerischen Heimat. Er wurde erst Leiter der Gütewirtschaft bei dem Unternehmen 'Deutsche Edelstahlwerke' in Krefeld, wechselte dann zum 'Verein Deutscher Eisenhüttenleute', wo er als Geschäftsführer des Bereichs „Technische Statistik“ bis 1968 wirkte. Nach zwei weiteren Jahren, in denen er Mitarbeiter des Vorstandes der 'Maximilianshütte' im bayerischen Sulzbach-Rosenberg war, folgte er im April 1970 dem Ruf an die Technische Universität Berlin auf den Lehrstuhl für Statistik und Wirtschaftsmathematik.

Die lange Tätigkeit in der Praxis prägten Bauers Forschungsinteressen nachhaltig. Etliche seiner über sechzig Veröffentlichungen sind der statistischen Qualitätskontrolle gewidmet. Gerade in seinen letzten wissenschaftlichen Arbeiten hat er sich mit diesem Problembereich beschäftigt. Auch blieb er aktives Mitglied in verschiedenen Gesellschaften wie dem VDEh und der EOQC, die sich mit der Qualitätskontrolle auseinandersetzen. Die Verbindungen zu seinen früheren Arbeitsstätten hat er immer aufrecht erhalten. Als Hochschullehrer spielte er sogar lange mit dem Gedanken, wieder in die Stahlindustrie zurückzugehen. Forschung (und Lehre, aber davon wird noch zu sprechen sein) konnte er sich nicht losgelöst von praktischem Bezug vorstellen.

Dabei war die Qualitätskontrolle gar nicht das Betätigungsfeld, das er seinen ersten Studenten und Assistenten zu Beginn der 70'er Jahre vorwiegend nahebrachte. Viel intensiver erfuhren sie seine Interessen in der Biometrie. Auch die Wurzeln dieses Schwerpunktes reichen weit vor den Antritt der Stelle an der TU zurück. Bereits als Student beschäftigte er sich mit anthropometrischen Fragen. Wie wach sein Interesse dafür die ganze Zeit blieb, zeigt sich unter anderem daran, daß er an fast allen Jahrestagungen der 'Deutschen Region der Biometric Society' teilnahm. Eine in diesem Zusammenhang erwähnenswerte, und über längere Zeit verfolgte Idee bestand in der Entwicklung einer am Menschen orientierten Gesamtdarstellung der statistischen Methoden, einer „Humanstatistik“.

Bauer war keineswegs auf diese Fragestellungen festgelegt, sondern neuen Problem-bereichen gegenüber sehr aufgeschlossen. Als die 'Deutsche Statistische Gesellschaft' beschloß, sich mit den Fragen der Umweltstatistik auseinanderzusetzen, hielt er 1974 auf der 45. Jahrestagung der Gesellschaft das Grundsatzreferat „Methodische Grundprobleme der Umweltstatistik“. In dem Referat kommt Bauer zu dem Fazit: „Es ist... die Beziehung herzustellen zwischen dem, was gemessen werden kann, und dem, was das Gemessene bewirkt“. Mit Worten aber begnügte er sich typischerweise nicht. Schon 1975 wurde das erste Projekt im Bereich Umweltstatistik initiiert. Die 'Berliner Hausmüllanalyse' bildete dann den Ausgangspunkt für die 'Arbeitsgruppe Umweltstatistik', in der zahlreiche praktische Forschungsprojekte bearbeitet wurden und die bis heute einen

festen Bestandteil des Institutslebens darstellt. Er hat vieles andere mehr angepackt und auch viele Fragen nun offen zurückgelassen.

In den letzten Jahren wandte sich sein Interesse immer mehr der Theorie der Paarvergleiche zu. Bekanntlich haben Kendall und Babington Smith in ihrer fundamentalen Arbeit als Inkonsistenzmaß die Anzahl T der zirkulären Tripel vorgeschlagen. Nun lassen sich offenbar auch andere Inkonsistenzmaße denken, als Alternativen oder Ergänzungen zu T . Die Fragestellung von R. K. Bauer war jetzt die folgende: wie läßt sich die anscheinend unendliche Vielfalt derartiger Maße ordnen? Eine zentrale Erkenntnis Bauers ist die Tatsache, daß es in Wirklichkeit nur endlich viele wesentlich verschiedener Maße gibt. Dazu definiert er einen geeigneten Begriff der Isomorphie zweier Präferenzschemata, der es erlaubt, die Menge aller solchen Schemata in Klassen „gleich inkonsistenter“ Schemata zu zerlegen. Innerhalb einer solchen Klasse wird nun ein vernünftiges Inkonsistenzmaß keine verschiedenen Werte annehmen. Die „Bauer-Isomorphie“ von gerichteten Graphen ist dabei ein sehr fruchtbares Konzept, welches zum üblichen Isomorphie-Begriff von gerichteten Graphen in vergleichbarer Beziehung steht wie die „2-Isomorphie“ zur gewöhnlichen Isomorphie bei den ungerichteten Graphen. Mit dem weiterführenden Gedanken, auf den durch „Bauer-Isomorphie“ gewonnenen Inkonsistenzklassen eine Partialordnungsrelation für die „Inkonsistenz“ einzuführen, erhält man dann eine vollständige Übersicht aller brauchbaren Inkonsistenzmaße. Diese Gedanken Bauers sind von großer Bedeutung für die Theorie der Paarvergleiche und es wäre sehr wünschenswert für den Fortschritt dieser Theorie, wenn seine Ansätze weitergetragen würden.

Bauer hat immer betont, daß er sich an der Universität in erster Linie als Hochschullehrer sehe. Dieses Selbstverständnis führte dazu, daß er sein Lehrdeputat stets übererfüllte. Augenfälliger noch zeigt es sich in der von ihm initiierten Organisation des Statistik- und Mathematik-Grundstudiums für Wirtschaftswissenschaftler an der TU in Kleingruppen. Wie sehr ihm die Ersetzung der großen Massenvorlesung durch diese Form der Wissensvermittlung am Herzen lag, macht die Erklärung deutlich, mit der er seine Bereitschaft zur Übernahme des Vorsitzes des Fachbereichsrates begründet. Hier nennt er ausdrücklich die diesbezüglich im Fachbereich erfahrene Unterstützung als entscheidenden Faktor.

Bei seinen Veranstaltungen im Hauptstudium konnte er den von ihm als wesentlich angesehenen Praxisbezug einbringen. Aber nicht nur damit und mit der noch lange nicht verbreiteten nichtparametrischen Ausrichtung verstand er es, seine Zuhörer für den zumeist nichtparametrisch aufbereiteten Stoff zu motivieren. Seinem Hobby, der Schauspielerei, verdankte er auch einige Kniffe, die er bei Gelegenheit nutzte. Aber das war wohl kaum der Grund, daß ihn seine Studenten so verehrten. Eher ist die Begründung schon darin zu suchen, daß sein Engagement für die Studenten lange nicht bei der Lehre aufhörte. Um ihre Belange kümmerte er sich auch in verschiedenen Gremien der akademischen Selbstverwaltung. So war er seit 1971 ohne Unterbrechung Vorsitzender der studentischen Darlehenskasse.

Bauer war in zahlreichen Selbstverwaltungsgremien aktiv. Gerade die, die er als Mitarbeiter der ersten Generation an Bauers Institut selbst in den Gremien arbeiteten, wissen insbesondere um seinen Einsatz als Vorsitzender des Fachbereichsrats, um widerstrebende Interessen beim Aufbau des neuen Fachbereichs Informatik zusammenzuführen. Er war natürlich in die Selbstverwaltung des Instituts und des Fachbereichs eingebunden, war aber weiter erster Vizepräsident und Mitglied im akademischen Senat der TU. Seinen Mitarbeitern hat er seiner demokratischen Grundeinstellung gemäß stets weitgehende Freiheiten gewährt und eigenen Aktivitätsdrang gefördert. So konnten die Assistenten eigenverantwortlich Lehrveranstaltungen im Grund- und Hauptstudium durchführen, eigene Konzepte einbringen und Projekte realisieren. Er war zwar ihr Vorgesetzter, doch das hat er sie nie so empfinden lassen.

Daß er für die Mitarbeiter stets ein offenes Ohr hatte, brachte ein gewaltiges Arbeitspensum mit sich. Er war an 15 Habilitationen und über 80 Promotionen beteiligt, und nicht nur für die zahlreichen Assistenten mit noch laufenden Promotionsverfahren kam sein Tod allzu früh. Wie wichtig ihm die Menschen, die anderen waren, zeigte sich auch bei seinem kürzlich 20jährigen Dienstjubiläum, wo er sich einerseits über das Erscheinen zahlreicher „Altgedienter“ freute, andererseits aber seinen Schmerz über zwei, die nicht mehr dabei sein konnten, kaum zu bewältigen vermochte.

Bei dem Einsatz für die Belange seiner Studenten und der Mitarbeiter an seinem Institut war Bauer stets ebenso geradlinig und engagiert wie bei seinen anderen gesellschaftlichen Handlungen. Dabei hat er sich immer von seiner konsequent demokratischen Haltung leiten lassen, was jedoch nie dazu führte, daß er eigene Maßstäbe aufgegeben hätte. Dies hat ihm auch bei kontroversen Positionen Respekt, ja Hochachtung eingebracht. Bekanntlich hat Bauer nach der turbulenten Mitgliederversammlung im März 1990 seinen Austritt aus der Deutschen Region der Biometrischen Gesellschaft erklärt, ein Schritt, der von vielen Mitgliedern sehr bedauert wurde. Bei allem Bedauern sollte dieser Entscheidung aber ebenfalls mit Respekt begegnet werden, zumal Bauer seinen Austritt mit dem Prinzip der wechselseitigen Loyalität von Lehrer und Schüler begründete. Ungeachtet seiner Entscheidung, die Region Deutschland zu verlassen und sich der Region Österreich-Schweiz (ROeS) unserer Gesellschaft anzuschließen, hatte sich Bauer sehr darauf gefreut, anlässlich des Biometrischen Kolloquiums in Hamburg wieder über neue Resultate zur Theorie der Paarvergleiche vortragen zu können.

Bei aller Bestätigung und Anerkennung, die er an seinen Wirkungsstätten erfuhr, blieb die Verbundenheit Bauers mit seiner bayerischen Heimat stets lebendig. Sie wurde von ihm gepflegt und gehörte zu seinem Wesen und zu seinem Erscheinungsbild, zu dem auch ein großes Emblem der Ostermarschierer beitrug. Höhepunkte, an denen die Liebe zu seiner Heimat deutlich wurden, bildeten die Faschingsvorlesungen, in denen er über seine Forschungen zu speziellen Fragen des Bairischen berichtete, wie etwa über 'Bairische Wortlehre: Artikel, Partikel und Interjektionen'.

Die Gemeinschaft der Statistiker und die Internationale Biometrische Gesellschaft hat mit Rainald K. Bauer einen markanten, bedeutenden Vertreter verloren. Seine farbige und menschliche Persönlichkeit wird uns fehlen.

Rainer Schlittgen
Bernd Streitberg

Nonparametric approaches to the analysis of multiple endpoints

James A. Koziol

Abstract

Nonparametric approaches to the analysis of multiple endpoints are described, and illustrated with data arising from a recent clinical trial of t-PA in acute stroke.

Key words:

multivariate rank statistics, permutation principle, clinical trials, case-control studies, paired comparisons, matching

I. Introduction

Treatment comparisons in randomized clinical trials can typically involve several different endpoints, which may in turn raise concerns for appropriate methods of data presentation and analysis. Strategies for addressing this multiple endpoint problem have been proposed previously, notably by O'BRIEN (1984) and POCOCK, GELLER, and TSIATIS (1987). The purpose of this note is to point out that the analysis of multiple endpoint data falls naturally within the province of multivariate rank statistics. In particular, if the endpoints are continuous, the multivariate distribution-free statistics introduced by KOZIOL et al. (1981, 1982) are immediately applicable; in addition, these statistics may be readily adapted to scenarios in which the multiple endpoints are mixed continuous and discrete, or solely discrete. These statistics are described, and examples are given in the following sections. The multivariate rank statistic approach is particularly attractive with small sample sizes or when the assumption of multivariate normality of the various endpoints may be untenable or unwarranted; this approach thus provides a useful complement to most of the strategies previously proposed.

II. Multivariate rank statistics: Application to the multiple endpoint problem

Let F_{ik} denote the marginal cumulative distribution function of the k^{th} variable for the i^{th} group, $i = 1, 2, \dots, I$, $k = 1, 2, \dots, K$, and let F_i denote the K -variate cumulative distribution function of the K variables for the i^{th} group. The distributions are identical under the null

hypothesis of no group differences, so the relevant hypothesis-testing problem may be formulated as a test of

$$H_0 : F_i = F \text{ for all } i = 1, 2, \dots, I, \quad (2.1)$$

against the general alternative that at least one inequality among the F_i obtains; here, F is an arbitrary element in \mathcal{F} , the set of all K -variate distributions.

To test the hypothesis (2.1), KOZIOL et al. (1981) devised a multivariate rank statistic by use of the basic rank permutation principle discussed in detail by CHATTERJEE and SEN (1964, 1966); see also PURI and SEN (1969, 1971). This approach entails first a joint ranking of all observations across the I groups on each of the K marginal variables, then construction of univariate rank statistics in the manner of HÁJEK and SÍDÁK (1967) for each of the K marginal variables. The rank permutation principle leads to determination of the joint multivariate distribution of the marginal univariate rank statistics.

Although KOZIOL et al. (1981) assumed that the F_i are continuous, it may well occur that certain of the F_{ik} may violate this assumption, so that ties may occur among the observations on the k^{th} variable. In this instance, we immediately apply the average score method (e.g., HÁJEK and SÍDÁK (1967); HÁJEK (1969)) for the assignment of scores to tied ranks: that is, each random variable in a tie is assigned the average of the scores that belong to all the random variables in that tie. Recall that extensions of univariate rank statistics to data arising from non-continuous distribution functions have been studied by VORLICKOVÁ (1970, 1972) and CONOVER (1973), among others. Hence, it should not be surprising that the basic rank permutation principle may be suitably modified to deal with tied data: in particular, following KOZIOL et al. (1981), KOZIOL (1990) established that the joint conditional asymptotic null distribution of the marginal linear rank statistics is multivariate normal.

KOZIOL et al. (1981) had exploited this fact to formulate two classes of test statistics with readily amenable approximating distributions for the hypothesis testing problem (2.1). The first class is patterned after Hotelling's T^2 test with normally distributed data in that it provides omnibus tests of no group differences. However, as MEIER (1975) has argued, the global alternative of any departure from the null hypothesis (2.1) is generally less useful in a clinical trials context than a more focussed alternative, in which one group is superior to the others as measured by the several endpoints.

The second class of statistics formulated by KOZIOL et al. (1981) is directed toward this stochastically ordered alternative that the group differences are in the same direction. This class is based on a quadratic form, denoted M_N by KOZIOL et al. (1981, eqn. 2.8). Under the permutation principle, suitably large values of M_N would lead to rejection of H_0 . In situations where determination of the exact permutation distribution of M_N is computationally prohibitive, asymptotic critical values are available upon reference to the appropriate chi-square distribution.

Let us make three remarks relevant to these classes of test statistics. (A) O'BRIEN's (1984) approach to the analysis of multiple endpoints is closely related to M_N : O'Brien utilizes Wilcoxon scores on each marginal, followed by summing across marginals as with M_N . However, O'Brien then suggests a oneway analysis of variance test procedure, in contrast to the more exact permutational distribution test advocated here. (B) In a similar spirit, we may improve upon the operating characteristics of M_N by incorporating suitable weights into the summary rank statistics on which it is based. However, precise knowledge of the likely alternative would be needed in order to optimize the power properties of the resultant tests [cf. WEI and JOHNSON (1985)]. (C) Note that in the situations of discrete or mixed continuous-discrete data, the permutation distributions of these test statistics may not necessarily be strictly distribution-free, as the patterns of ties are reflective of the underlying distributions. Nevertheless, the conditional asymptotic distributions of the test statistics are asymptotically distribution-free in the sense of LEHMANN (1975).

We illustrate the multivariate rank statistic approach with the following example. Incorporated into human plasma is an enzymatic system for dissolving the fibrin in blood clots. One component of this system is tissue plasminogen activator (t-PA). t-PA mediates the proteolysis of plasminogen to form the active enzyme plasmin, which in turn degrades the fibrin network of a clot to form soluble products. It is therefore a potentially useful thrombolytic agent for the treatment of pulmonary embolisms, deep-vein thromboses, heart attacks and strokes. In this regard, a pilot study was recently conducted at the Scripps Clinic and Research Foundation to investigate its efficacy in the treatment of stroke. In particular, the purpose of the study was to test the hypothesis that thrombus-directed fibrinolysis by t-PA at a fixed dose may lead to neurologic recovery, when given early (within 6 hours from symptom onset) by intravenous infusion to patients presenting with acute stroke. A total of 12 patients who met the study entry criteria were randomly assigned either to the control arm or to the intravenous t-PA arm (6 patients to each). Outcomes were assessed according to two criteria: (i) the extent of recanalization of the occluded artery (ies), as determined by cerebral angiography; and (ii) the improvement in neurological status. [All patients were essentially equivalent in terms of both perfusion status of their respective affected arterial territories, and neurological status, at time of presentation.]

We proceed with Wilcoxon-type rank scores for each of the outcome variates for purposes of illustration; the relevant data are given in Table 1. One may readily verify that marginally the ranked data are suggestive of t-PA efficacy, though the evidence is not overwhelming. Note that here with the small sample sizes involved, the determination of the exact conditional permutation distributions of these various statistics is a computationally feasible task. In principle, with such exactly determined achieved significance levels, the concern over Type I error rate probabilities is obviated. The overall statistic M_N against ordered alternatives reduces here to the sum of the Wilcoxon scores on each outcome parameter for either group, and, as noted previously, is thus somewhat analogous to O'BRIEN's (1984) nonparametric procedure for multiple endpoints. We found that M_N achieves a (one-sided) exact permutation significance level of 0.05 in testing (2.1) against the ordered alternative that t-PA improves both clinical outcomes; this may be compared with an estimated significance level of .04 from the normal approximation to the permutation distribution of M_N . In sum, combining information can add to the weight of evidence in support of treatment differences in clinical trials. In this particular instance, the pilot study has indicated the feasibility and desirability of a larger, more comprehensive clinical trial of t-PA in treating acute stroke.

We conclude with the remark that KOZIOL and MAXWELL (1982) have provided an extension of the multivariate rank statistic methodology of KOZIOL et al. (1981) to multivariate signed rank statistics. Koziol and Maxwell had examined paired growth curves with this approach; but the approach is immediately applicable for assessing the equivalence of paired multivariate measurements, as would arise in paired or matched case-control studies with multiple

Table 1: Ranks of Angiography and Neurological Status Outcomes for the Two Groups

Angio	Control Arm		Angio	t-PA Arm	
		Neuro			Neuro
3		6.5	1		3
5		3	2		1
8		9.5	4		3
10		11	6		6.5
11		6.5	7		9.5
12		12	9		6.5

endpoints. Since details of the Koziol-Maxwell approach are readily available, we here dispense with theoretical concerns, and refer the reader to KOZIOL (1990) for an illustrative example.

References

1. CHATTERJEE, S. K. and SEN, P. K. Non-parametric tests for the bivariate two sample location problem. *Calcutta Statistical Association Bulletin* 13 : 18-58, 1964.
2. CHATTERJEE, S. K. and SEN, P. K. Nonparametric tests for the multisample multivariate location problem. In: *Essays in Probability and Statistics in Memory of S. N. Roy, R. C. Bose et al.* Eds. Chapel Hill, University of North Carolina Press, 1966, pp. 197 – 228.
3. CONOVER, W. J. Rank tests for one sample, two samples, and k samples without the assumption of a continuous distribution function. *Ann. Stat.* 1 : 1105 – 1125, 1973.
4. HÁJEK, J. and SÍDÁK, Z. *Theory of Rank Tests*, Academic Press, New York, 1967.
5. HÁJEK, J. *A Course in Nonparametric Statistics*, Holden-Day, San Francisco, 1969.
6. KOZIOL, J. A. Nonparametric approaches to the analysis of multiple endpoints. Unpublished technical report, 1990.
7. KOZIOL, J. A. and MAXWELL, D. A. A distribution-free test for paired growth curve analyses with application to an animal tumor immunotherapy experiment. *Stat. Med.* 1 : 83 – 89, 1982.
8. KOZIOL, J. A. MAXWELL, D. A., FUKUSHIMA, M., COLMERAUER, M. E. M. and PILCH, Y. H. A distribution-free test for tumor growth curve analyses with application to an animal tumor immunotherapy experiment. *Biometrics* 37 : 383 – 390, 1981.
9. LEHMACHER, W. WASSMER, G. and REITMEIR, P. Procedures for two-sample comparisons with multiple endpoints controlling the experimentwise error rate. To appear in *Biometrics*, 1991.
10. LEHMANN, E. L. *Nonparametrics: Statistical Methods Based on Ranks*, Holden-Day, San Francisco, 1975.
11. MEIER, P. Statistics and medical experimentation. *Biometrics* 31 : 511 – 529, 1975.
12. O'BRIEN, P. C. Procedures for comparing samples with multiple endpoints. *Biometrics* 40 : 1079 – 1087, 1984.
13. POCOCK, S. J., GELLER, N. L. AND TSIATIS, A. A. The analysis of multiple endpoints in clinical trials. *Biometrics* 43 : 487 – 498, 1987.
14. PURI, M. L. and SEN, P. K. A class of rank order tests for a general linear hypothesis. *Ann. Math. Stat.* 40 : 1325 – 1343, 1969.
15. PURI, M. L. and SEN, P. K. *Nonparametric Methods in Multivariate Analysis*, John Wiley, New York, 1971.
16. SÍDÁK, Z. Rectangular confidence regions for the means of multivariate normal distributions. *J. Am. Stat. Assoc.* 62 : 626 – 633, 1967.
17. VORLICKOVÁ, D. Asymptotic properties of rank tests under discrete distributions. *Z. Wahrscheinlichkeitstheorie und verw. Gebiete* 14 : 275 – 289, 1970.
18. VORLICKOVÁ, D. Asymptotic properties of rank tests of symmetry under discrete distributions. *Ann. Math. Stat.* 43 : 2013 – 2018, 1972.
19. WEI, L. J. and JOHNSON, W. F. Combining dependent tests with incomplete repeated measurements. *Biometrika* 72 : 359 – 364, 1985.

Planung und Auswertung der dreifaktoriellen Streifenspaltanlage (A+B)/C-BI und der Spaltstreifenanlage A/(B+C)-BI

H. Dörfel und T. Bauer

Zusammenfassung

Es werden Planung und Auswertung der dreifaktoriellen Streifenanlage (A+B)/C-BI und der dreifaktoriellen Spaltstreifenanlage A/(B+C)-BI dargestellt. Die Formeln für die Schätzung der Varianz der Mittelwertdifferenzen für alle 19 denkbaren Mittelwertvergleiche werden für beide Versuchsanlagen angegeben.

Summary

Design and analysis of the Strip-Split-Pilot-Design (A+B)/C-BI and of the Split-Strip-Plot-Design A/(B+C)-BI both including three factors are described. The formula for the estimation of the standard error of difference between means are given for all 19 comparisons between means for both designs.

1. Einleitung

In der Literatur sind die Planung und Auswertung von dreifaktoriellen Versuchsanlagen mit eingeschränkter Randomisation der Prüfglieder in vollständigen Blocks beschrieben (COCHRAN u. COX, 1957; MUDRA, 1958; THOMAS, 1966a, b; DÖRFEL u. BÄTZ, 1973; FEDERER, 1975; Autorenkollektiv, 1987). Solche Versuchsanlagen werden im Feldversuchswesen verwendet, wenn die agrotechnische Durchführung der Stufen der Prüffaktoren keine vollständige Randomisation der Prüfglieder im Block gestattet.

In dieser Arbeit werden zwei Versuchsanlagen vorgestellt, deren varianzanalytische Auswertung, aber besonders die Varianzen für die verschiedenen Mittelwertvergleiche, die bisher in der Literatur nicht veröffentlicht sind.

Die Ableitung der Varianzen für alle denkbaren Mittelwertvergleiche in einem dreifaktoriellen Versuch erfolgt für die Versuchsanlagen auf analoge Weise zur Arbeit von DÖRFEL u. BÄTZ (1973).

2. Beschreibung der Versuchsanlagen

Für die Konstruktion der Streifenspaltanlage (A+B)/C-BI bildet eine zweifaktorielle Streifenanlage A+B-BI den Ausgangspunkt. Die Symbole für die Versuchsanlagen werden in Abschnitt 3 erläutert. In jedem „Teilstück“ dieser Streifenanlage werden die Stu-

fen des Prüffaktors C zufällig angeordnet. In Abbildung 1 ist ein Block dieser Streifenspaltanlage dargestellt. Es wird empfohlen, die zufällige Anordnung der Stufen der Prüffaktoren A und B von Block zu Block und die zufällige Anordnung der Stufen des Prüffaktors C von „AB-Teilstück“ zu „AB-Teilstück“ zu verändern.

Für die Konstruktion der Spaltstreifenanlage $A/(B+C)$ -Bl bilden eine einfaktorielle Blockanlage A-Bl mit a Stufen und r Wiederholungen sowie eine zweifaktorielle Streifenanlage $B+C$ -Bl mit $a \cdot r$ Wiederholungen den Ausgangspunkt. In jedem „Teilstück“ der einfaktoriellen Blockanlage wird eine „Wiederholung“ der zweifaktoriellen Streifenanlage angeordnet. In Abbildung 2 ist ein Block dieser Spaltstreifenanlage dargestellt. Es wird empfohlen, die zufällige Anordnung der Stufen des Prüffaktors A von Block zu Block und die zufällige Anordnung der Stufen der Prüffaktoren B und C von „A-Teilstück“ zu „A-Teilstück“ zu verändern.

	a1			a2		
b1	alb1c3	alb1c1	alb1c2	a2b1c1	a2b1c3	a2b1c2
b2	alb2c2	alb2c3	alb2c1	a2b2c3	a2b2c2	a2b2c1

Abbildung 1: Block der dreifaktoriellen Streifenspaltanlage $(A+B)/C$ -Bl mit $a = 2$, $b = 2$, $c = 3$

	a1			a2		
b1	alb1c3	alb1c1	alb1c2	a2b2c1	a2b2c3	a2b2c2
b2	alb2c3	alb2c1	alb2c2	a2b1c1	a2b1c3	a2b1c2
	c3	c1	c2	c1	c3	c2

Abbildung 2: Block der dreifaktoriellen Spaltstreifenanlage $A/(B+C)$ -Bl mit $a = 2$, $b = 2$, $c = 3$

3. Symbole der Versuchsanlagen

Die Symbole zur Kurzbezeichnung der Versuchsanlagen mit vollständigen Blocks sind in einem Fachbereichsstandard (TGL 21168/14) der Akademie der Landwirtschaftswissenschaften der ehemaligen DDR festgelegt. Da diese Symbolik aus unserer Sicht empfehlenswert ist, soll sie hier kurz beschrieben werden.

Die Symbole bestehen aus zwei Teilen, die durch einen Bindestrich „-“ miteinander verbunden sind.

Der Teil nach dem Bindestrich gibt an, ob es sich um eine Versuchsanlage

- ohne Blockbildung: -R
- mit einfacher Blockbildung: -Bl
- mit zweifach orthogonaler Blockbildung
 - Lateinisches Rechteck: -LR
 - Lateinisches Quadrat: -LQ

handelt.

Die Anzahl der Großbuchstaben vor dem Bindestrich ist gleich der Anzahl der Prüffaktoren im Versuch.

In Versuchen mit zwei oder mehr Prüffaktoren sind die Großbuchstaben durch verschiedene Zeichen voneinander getrennt.

Aus diesen Zeichen geht die Zuordnung der Stufen der Prüffaktoren zu den Teilstücken

eines Blocks hervor. An zweifaktoriellen Versuchen mit einfacher Blockbildung wird die Symbolik dargestellt. Dabei bedeutet

- a = Anzahl der Stufen des Prüffaktors A und
- b = Anzahl der Stufen des Prüffaktors B.

AxB-BI ist das Symbol für eine zweifaktorielle Blockanlage. Das Zeichen „x“ bedeutet, daß jeder Block in a*b Teilstücke unterteilt ist. Diesen werden die Stufen beider Prüffaktoren unabhängig voneinander zufällig zugeordnet.

A/B-BI ist das Symbol für eine zweifaktorielle Spaltanlage. Das Zeichen „/“ bedeutet, daß jeder Block in a Großteilstücke und jedes dieser in b Kleinteilstücke unterteilt ist. Die Stufen des Prüffaktors A werden den Großteilstücken und die Stufen des Prüffaktors B den Kleinteilstücken in jedem Großteilstück zufällig zugeordnet.

A+B-BI ist das Symbol für eine zweifaktorielle Streifenanlage. Das Zeichen „+“ bedeutet, daß jeder Block in a Zeilen und b Spalten unterteilt ist. Die Stufen des Prüffaktors A werden den Zeilen und die Stufen des Prüffaktors B den Spalten zufällig zugeordnet.

4. Dreifaktorielle Streifenspaltanlage (A+B)/C-BI

4.1. Modell

Für die Beobachtungswerte y_{ijkl}

i = 1...a Stufen des Prüffaktors A

j = 1...b Stufen des Prüffaktors B

k = 1...c Stufen des Prüffaktors C

l = 1...r Anzahl der Wiederholungen (Blocks)

gilt das Modell

$$y_{ijkl} = \mu + a_i + b_j + c_k + (ab)_{ij} + (ac)_{ik} + (bc)_{jk} + (abc)_{ijk} + \underline{w}_l + \underline{e}_{il} + \underline{e}_{jl} + \underline{e}_{kl} + \underline{e}_{ijkl} \quad (1)$$

Für die festen Effekte der Stufen der Prüffaktoren und deren Wechselwirkungseffekte gelten die bekannten Reparametrisierungsbedingungen.

$$a_i \quad \text{Effekt der i-ten Stufe von A (i=1...a): } \sum a_i = 0$$

$$b_j \quad \text{Effekt der j-ten Stufe von B (j=1...b): } \sum b_j = 0$$

$$c_k \quad \text{Effekt der k-ten Stufe von C (k=1...c): } \sum c_k = 0$$

(ab)_{ij} Wechselwirkungseffekt der i-ten A- und der j-ten

$$\text{B-Stufe: } \sum_i (ab)_{ij} = \sum_j (ab)_{ij} = 0$$

(ac)_{ik} Wechselwirkungseffekt der i-ten A- und der k-ten

$$\text{C-Stufe: } \sum_i (ac)_{ik} = \sum_k (ac)_{ik} = 0$$

(bc)_{jk} Wechselwirkungseffekt der j-ten B- und der k-ten

$$\text{C-Stufe: } \sum_j (bc)_{jk} = \sum_k (bc)_{jk} = 0$$

$(abc)_{ijk}$ Wechselwirkungseffekt der i-ten A-, j-ten B- und k-ten

$$\text{C-Stufe: } \sum_i (abc)_{ijk} = \sum_j (abc)_{ijk} = \sum_k (abc)_{ijk} = 0$$

Die w_l , e_{il} , e_{jl} und e_{ijkl} sind normalverteilte unabhängige Zufallsvariable mit dem Erwartungswert Null und den Varianzen

$$\begin{aligned} \text{Var}(w_l) &= \sigma^2_{Bl} && \text{für alle } l \\ \text{Var}(e_{il}) &= \sigma^2_{RA} && \text{für alle } i, l \\ \text{Var}(e_{jl}) &= \sigma^2_{RB} && \text{für alle } j, l \\ \text{Var}(e_{ijl}) &= \sigma^2_{RAB} && \text{für alle } i, j, l \\ \text{Var}(e_{ijkl}) &= \sigma^2_{RABC} && \text{für alle } i, j, k, l \end{aligned}$$

4.2. Varianzanalyse

Entsprechend Modell (1) kann die Varianzanalyse durchgeführt werden. In Tabelle 1 ist die Varianztabelle mit den Schätzformeln für die verschiedenen Summen der Abweichungsquadrate (Rest) und deren Freiheitsgrade dargestellt.

Tabelle 1: Varianztabelle und Schätzformeln für die verschiedenen SQ(Rest)

Ursache	SQ	FG
Gesamt		
A+Block	$bc \sum_i \sum_l (\bar{y}_{i..l} - \bar{y}....)^2$	ar-1
Block		r-1
A		a-1
Rest A	$SQ(A+Bl) - SQ(Bl) - SQ(A)$	(a-1)(r-1)
B+Block	$ac \sum_j \sum_l (\bar{y}_{j.l} - \bar{y}....)^2$	br-1
Block		r-1
B		b-1
Rest B	$SQ(B+Bl) - SQ(Bl) - SQ(B)$	(b-1)(r-1)
A+B+Block	$c \sum_i \sum_j \sum_l (\bar{y}_{ij.l} - \bar{y}....)^2$	abr-1
A x B		(a-1)(b-1)
Rest AB	$SQ(A+B+Bl) - (SQ(Bl) - SQ(A+B) - SQ(RA) - SQ(RB))$	(a-1)(b-1)(r-1)
C		c-1
A x C		(a-1)(c-1)
B x C		(b-1)(c-1)
A x B x C		(a-1)(b-1)(c-1)
Rest ABC	$SQ(\text{Ges.}) - SQ(Bl) - SQ(A+B+C) - SQ(RA) - SQ(RB) - SQ(RAB)$	ba(c-1)(r-1)

Die durchzuführenden F-Tests zeigt die nachfolgende Aufstellung.

	Nenner			
	MQ(RA)	MQ(RB)	MQ(RAB)	MQ(RABC)
MQ(A)	+			
MQ(B)		+		
MQ(C)				+
MQ(AxB)			+	
MQ(AxC)				+
MQ(BxC)				+
MQ(AxBxC)				+

4.3. Varianzen für die verschiedenen Mittelwertvergleiche

Neben den Vergleichen zwischen beliebigen (1) A-, (2) B-, (3) C-, (4) AB-, (5) AC-, (6) BC- und (7) ABC-Mitteln sind noch folgende Mittelwertvergleiche möglich:

Mittelwerte	auf gleicher Stufe von					
	A	B	C	AB	AC	BC
AB	(8)	(9)				
AC	(10)		(11)			
BC		(12)	(13)			
ABC	(14)	(15)	(16)	(17)	(18)	(19)

Die Schätzformeln der Varianzen der Mittelwertdifferenzen aller 19 denkbaren Mittelwertvergleiche sind in Tabelle 3 enthalten. Sie lassen sich aus den in Tabelle 2 enthaltenen Erwartungswerten der Restvarianzen ableiten.

Tabelle 2: Erwartungswerte der verschiedenen Restvarianzen

Rest A	MQ(RA)	$\sigma^2_{RABC} + c * \sigma^2_{RAB} + bc * \sigma^2_{RA}$
Rest B	MQ(RB)	$\sigma^2_{RABC} + c * \sigma^2_{RAB} + ac * \sigma^2_{RB}$
Rest AB	MQ(RAB)	$\sigma^2_{RABC} + c * \sigma^2_{RAB}$
Rest ABC	MQ(RABC)	σ^2_{RABC}

Für den Mittelwertvergleich (8) – AB-Mittelwerte auf gleicher A-Stufe – wird die Schätzformel für die Varianz der Mittelwertdifferenz hergeleitet:

$$\begin{aligned}
 \bar{y}_{ij..} - \bar{y}_{ij'..} &= (b_j - b_{j'}) + ((ab)_{ij} - (ab)_{ij'}) \\
 &\quad + \frac{1}{r} \sum (\epsilon_{jl} - \epsilon_{j'l}) \\
 &\quad + \frac{1}{r} \sum (\epsilon_{ijl} - \epsilon_{ij'l}) \\
 &\quad + \frac{1}{cr} \sum (\epsilon_{ijk} - \epsilon_{ij'k}) \\
 \text{Var}(\bar{y}_{ij..} - \bar{y}_{ij'..}) &= \frac{2}{r} * \sigma^2_{RB} + \frac{2}{r} * \sigma^2_{RAB} + \frac{2}{cr} * \sigma^2_{RABC} \\
 s^2_d(AB/A) &= \frac{2}{acr} \left(\text{MQ(RB)} + (a-1) \text{MQ(RAB)} \right)
 \end{aligned}$$

Tabelle 3: Varianzen der verschiedenen Mittelwertdifferenzen

Mittelwertvergleiche	Schätzformel der Varianz der Mittelwertdifferenz
(1) $\bar{y}_{i...} - \bar{y}_{i'...}$	$\frac{2}{bcr} * MQ(RA)$
(2) $\bar{y}_{.j..} - \bar{y}_{.j'..}$	$\frac{2}{acr} * MQ(RB)$
(3) $\bar{y}_{..k.} - \bar{y}_{..k'}$	$\frac{2}{abr} * MQ(RABC)$
(4) $\bar{y}_{ij..} - \bar{y}_{i'j'..}$	$\frac{2}{abcr} * (aMQ(RA) + MQ(RB) + (ab-a-b)MQ(RAB))$
(8) $\bar{y}_{ij..} - \bar{y}_{ij'..}$	$\frac{2}{acr} * (MQ(RB) + (a-1)MQ(RAB))$
(9) $\bar{y}_{ij..} - \bar{y}_{i'j'..}$	$\frac{2}{bcr} * (MQ(RA) + (b-1)MQ(RAB))$
(5) $\bar{y}_{i.k.} - \bar{y}_{i'.k'}$	$\frac{2}{bcr} * (MQ(RA) + (c-1)MQ(RABC))$
(11) $\bar{y}_{i.k.} - \bar{y}_{i'.k.}$	siehe Vergleich (5)
(10) $\bar{y}_{i.k.} - \bar{y}_{i'.k'}$	$\frac{2}{br} * MQ(RABC)$
(6) $\bar{y}_{.jk.} - \bar{y}_{.j'.k'}$	$\frac{2}{acr} * (MQ(RB) + (c-1)MQ(RABC))$
(13) $\bar{y}_{.jk.} - \bar{y}_{.j'.k.}$	siehe Vergleich (6)
(12) $\bar{y}_{.jk.} - \bar{y}_{.j'.k'}$	$\frac{2}{ar} * MQ(RABC)$
(7) $\bar{y}_{ijk.} - \bar{y}_{i'j'k'}$	$\frac{2}{abcr} * (aMQ(RA) + bMQ(RAB) + (ab-a-b)MQ(RAB) + (abc-ab)MQ(RABC))$
(16) $\bar{y}_{ijk.} - \bar{y}_{i'j'k.}$	siehe Vergleich (7)
(14) $\bar{y}_{ijk.} - \bar{y}_{ij'k'}$	$\frac{2}{acr} * (MQ(RB) + (a-1)MQ(RAB) + (ac-c)MQ(RABC))$
(18) $\bar{y}_{ijk.} - \bar{y}_{ij'k.}$	siehe Vergleich (14)
(15) $\bar{y}_{ijk.} - \bar{y}_{i'j'k'}$	$\frac{2}{bcr} * (MQ(RA) + (b-1)MQ(RAB) + (bc-b)MQ(RABC))$
(19) $\bar{y}_{ijk.} - \bar{y}_{i'j'k.}$	siehe Vergleich (15)
(17) $\bar{y}_{ijk.} - \bar{y}_{ijk'}$	$\frac{2}{r} * MQ(RABC)$

5. Dreifaktorielle Spaltstreifenanlage A/(B+C)-BI

5.1. Modell

Für die Beobachtungswerte y_{ijkl} gilt das Modell

$$y_{ijkl} = \mu + a_i + b_j + c_k + (ab)_{ij} + (ac)_{ik} + (bc)_{jk} + (abc)_{ijk} + \underline{w}_l + \underline{e}_{il} + \underline{e}_{jl} + \underline{e}_{kl} + \underline{e}_{ijkl} \quad (2)$$

Für die festen Effekte der Stufen der Prüffaktoren und deren Wechselwirkungseffekte gelten die bekannten Reparametrisierungsbedingungen (siehe Abschnitt 4.1.). Die w_l , e_{il} , e_{ikl} und e_{ijkl} sind normalverteilte unabhängige Zufallsvariable mit dem Erwartungswert Null und den Varianzen

$$\begin{aligned} \text{Var}(w_l) &= \sigma^2_{Bl} && \text{für alle } l \\ \text{Var}(e_{il}) &= \sigma^2_{RA} && \text{für alle } i, l \\ \text{Var}(e_{ijl}) &= \sigma^2_{RAB} && \text{für alle } i, j, l \\ \text{Var}(e_{ikl}) &= \sigma^2_{RAC} && \text{für alle } i, k, l \\ \text{Var}(e_{ijkl}) &= \sigma^2_{RABC} && \text{für alle } i, j, k, l \end{aligned}$$

5.2. Varianzanalyse

Entsprechend Modell (2) kann die Varianzanalyse durchgeführt werden. In Tabelle 4 ist die Varianzabelle mit den Schätzformeln für die verschiedenen Summen der Abweichungsquadrate (Rest) und deren Freiheitsgrade dargestellt.

Tabelle 4: Varianztabelle und Schätzformeln für die verschiedenen SQ(Rest)

Ursache	SQ	FG
Gesamt		
A+Block	$bc \sum_i \sum_l (\bar{y}_{i..l} - \bar{y}_{....})^2$	ar-1
Block		r-1
A		a-1
Rest A	$SQ(A+Bl) - SQ(Bl) - SQ(A)$	(a-1)(r-1)
A+B+Block	$c \sum_i \sum_j \sum_l (\bar{y}_{ij..l} - \bar{y}_{....})^2$	abr-1
B		b-1
AxB		(a-1)(b-1)
Rest AB	$SQ(A+B+Bl) - SQ(A+B) - SQ(Bl) - SQ(RA)$	a(b-1)(r-1)
A+C+Bl	$b \sum_i \sum_k \sum_l (\bar{y}_{i..kl} - \bar{y}_{....})^2$	acr-1
C		c-1
AxC		(a-1)(c-1)
Rest AC	$SQ(A+C+Bl) - SQ(A+C) - SQ(Bl) - SQ(RA)$	a(c-1)(r-1)
BxC		(b-1)(c-1)
AxBxC		(a-1)(b-1)(c-1)
Rest ABC	$SQ(\text{Ges.}) - SQ(A+B+C) - SQ(Bl) - SQ(RA) - SQ(RAB) - SQ(RAC)$	a(b-1)(c-1)(r-1)

Die durchzuführenden F-Tests zeigt die nachfolgende Aufstellung.

Zähler	Nenner			
	MQ(RA)	MQ(RAB)	MQ(RAC)	MQ(RABC)
MQ(A)	+			
MQ(B)		+		
MQ(C)			+	
MQ(AxB)		+		
MQ(AxC)			+	
MQ(BxC)				+
MQ(AxBxC)				+

Die Schätzformeln der Varianzen der Mittelwertdifferenzen aller 19 denkbaren Mittelwertvergleiche (siehe Abschn. 3.3) sind in Tabelle 6 auf Seite 56 enthalten. Sie lassen sich aus den in Tabelle 5 enthaltenen Erwartungswerten der Restvarianzen ableiten.

Tabelle 5: Erwartungswerte der verschiedenen Restvarianzen

Rest	MQ	E(MQ)
Rest A	MQ(RA)	$\sigma^2_{RABC} + c * \sigma^2_{RAB} + b * \sigma^2_{RAC} + bc * \sigma^2_{RA}$
Rest AB	MQ(RAB)	$\sigma^2_{RABC} + c * \sigma^2_{RAB}$
Rest AC	MQ(RAC)	$\sigma^2_{RABC} + b * \sigma^2_{RAC}$
Rest ABC	MQ(RABC)	σ^2_{RABC}

Für den Mittelwertvergleich (14) – ABC-Mittelwerte auf gleicher A-Stufe – wird die Schätzformel für die Varianz der Mittelwertdifferenz hergeleitet:

$$\begin{aligned}
 \bar{y}_{ijk.} - \bar{y}_{ij'k'} &= (b_j - b_{j'}) + (c_k - c_{k'}) + ((ab)_{ij} - (ab)_{ij'}) \\
 &\quad + ((ac)_{ik} - (ac)_{ik'}) + ((bc)_{jk} - (bc)_{j'k'}) \\
 &\quad + \frac{1}{r} \sum (\underline{e}_{ijl} - \underline{e}_{ij'l}) + \frac{1}{r} \sum (\underline{e}_{ikl} - \underline{e}_{ik'l}) \\
 &\quad + \frac{1}{r} \sum (\underline{e}_{ijk} - \underline{e}_{ij'k'}) \\
 \text{Var}(\bar{y}_{ijk.} - \bar{y}_{ij'k'}) &= \frac{2}{r} * \sigma^2_{RAB} + \frac{2}{r} * \sigma^2_{RAC} + \frac{2}{r} * \sigma^2_{RABC} \\
 s^2_d(ABC/A) &= \frac{2}{bcr} (bMQ(RAB) + cMQ(RAC) + (bc-b-c)MQ(RABC))
 \end{aligned}$$

6. Zur Anwendung

Die beiden vorgestellten Versuchsanlagen stellen Versuchsanlagen mit eingeschränkter Randomisation der Prüfglieder in den Blocks dar. Solche Versuchsanlagen werden im Feldversuchswesen dann verwendet, wenn die vollständige Randomisation der Prüfglieder in den Blocks nicht möglich ist.

Tabelle 6: Varianzen der verschiedenen Mittelwertdifferenzen

Mittelwertvergleich	Schätzformel der Varianz der Mittelwertdifferenz
(1) $\bar{y}_{i...} - \bar{y}_{i'...}$	$\frac{2}{bcr} * MQ(RA)$
(2) $\bar{y}_{.j..} - \bar{y}_{.j'..}$	$\frac{2}{acr} * MQ(RAB)$
(3) $\bar{y}_{..k.} - \bar{y}_{..k'}$	$\frac{2}{abr} * MQ(RAC)$
(4) $\bar{y}_{ij..} - \bar{y}_{i'j'..}$	$\frac{2}{bcr} * (MQ(RA) + (b-1) MQ(RAB))$
(9) $\bar{y}_{ij..} - \bar{y}_{i'j..}$	siehe Vergleich (4)
(8) $\bar{y}_{ij..} - \bar{y}_{ij'..}$	$\frac{2}{cr} * MQ(RAB)$
(5) $\bar{y}_{i.k.} - \bar{y}_{i'.k'}$	$\frac{2}{bcr} * (MQ(RA) + (c-1) MQ(RAC))$
(11) $\bar{y}_{i.k.} - \bar{y}_{i'.k}$	siehe Vergleich (5)
(10) $\bar{y}_{i.k.} - \bar{y}_{i.k'}$	$\frac{2}{br} * MQ(RAC)$
(6) $\bar{y}_{.jk.} - \bar{y}_{.j'k'}$	$\frac{2}{abcr} * (bMQ(RAB) + cMQ(RAC) + (bc-b-c)MQ(RABC))$
(12) $\bar{y}_{.jk.} - \bar{y}_{.jk'}$	$\frac{2}{abr} * (MQ(RAC) + (b-1)MQ(RABC))$
(13) $\bar{y}_{.jk.} - \bar{y}_{.j'k}$	$\frac{2}{acr} * (MQ(RAB) + (c-1)MQ(RABC))$
(7) $\bar{y}_{ijk.} - \bar{y}_{i'j'k'}$	$\frac{2}{bcr} * (MQ(RA) + (b-1)MQ(RAB) + (c-1)MQ(RAC) + (b-1)(c-1) MQ(RABC))$
(15) $\bar{y}_{ijk.} - \bar{y}_{i'jk'}$	siehe Vergleich (7)
(16) $\bar{y}_{ijk.} - \bar{y}_{i'j'k}$	siehe Vergleich (7)
(19) $\bar{y}_{ijk.} - \bar{y}_{i'jk}$	siehe Vergleich (7)
(14) $\bar{y}_{ijk.} - \bar{y}_{ij'k'}$	$\frac{2}{bcr} * (bMQ(RAB) + cMQ(RAC) + (bc-b-c)MQ(RABC))$
(17) $\bar{y}_{ijk.} - \bar{y}_{ijk'}$	$\frac{2}{br} * (MQ(RAC) + (b-1)MQ(RABC))$
(18) $\bar{y}_{ijk.} - \bar{y}_{ij'k}$	$\frac{2}{cr} * (MQ(RAB) + (c-1)MQ(RABC))$

- Aus versuchstechnischen Gründen kann eine Streifenspaltanlage (A+B)/C-BI dann sinnvoll sein, wenn in einem Versuch z.B. die Prüffaktoren A: Bodenbearbeitung, B: Beregnung und C: Sorten sind und die Prüffaktoren A, B jeweils unterschiedliche Großteilstücke erfordern.

- Wenn zur Vermeidung der gegenseitigen Beeinflussung der Wirkung der Prüffaktoren breite Nachbarstreifen um jede Parzelle angelegt werden müssen, steigt der Flächenbedarf schnell an. Durch Verzicht auf die vollständige Randomisation der Prüfglieder im Block kann die notwendige Fläche je Block verringert werden. Aus diesen genannten Gründen wurde in Bad Lauchstädt in der Nähe von Halle/S. ein Dauerversuch mit den drei Prüffaktoren Beregnung ($a = 2$ Stufen), organische Düngung ($b = 5$ Stufen) und mineralische Düngung ($c = 5$ Stufen) als Spaltstreifenanlage $A/(B+C)$ -Bl angelegt.

Beide Versuchsanlagen $(A+B)/C$ -Bl und $A/(B+C)$ -Bl haben gegenüber einer vollständigen Blockanlage den bekannten Vorteil, daß sie die Versuchsdurchführung erheblich erleichtern und außerdem noch eine Verminderung der benötigten Versuchsfläche (Einsparung von Nachbarstreifen!) und der damit verbundenen Kosten bewirken.

Allerdings haben sie neben der eingeschränkten Randomisation einen weiteren großen Nachteil gegenüber der vollständigen Blockanlage, der darin besteht, daß die Anzahl der Freiheitsgrade für die Schätzung der Restvarianzen in der Regel sehr gering ist. Das führt dazu, daß diese Schätzungen sehr unsicher sind. Außerdem führt die geringe Freiheitsgradanzahl im F-Test der Varianzanalyse dazu, daß nur relativ große Haupt- und Wechselwirkungen erkannt werden und die Ablehnung der Alternativhypothese recht unsicher ist. Das Erkennen bzw. Nichterkennen von Wechselwirkungen ist aber von großer Bedeutung für die sich an die Varianzanalyse anschließenden Mittelwertvergleiche (DÖRFEL u. BÄTZ 1980). Die gleiche Unsicherheit hinsichtlich des Erkennens bzw. Nichterkennens wie beim F-Test besteht dann auch bei den Mittelwertvergleichen.

Aus den genannten Gründen sollte ein Versuchsansteller vor der Verwendung einer Versuchsanlage mit einer eingeschränkter Randomisation der Prüfglieder in den Blocks die Vorteile dieser Versuchsanlagen immer ihren Nachteilen gegenüberstellen.

Das Problem der Schätzung fehlender Werte in solchen Anlagen wird in einer späteren Arbeit dargestellt.

Die Auswertung einer Versuchsserie erfolgt über die Prüfgliedmittelwerte und den gepoolten Fehler für diese Mittelwerte (Formel (7) in Tabelle 3 bzw. 6).

7. Literatur

- Autorenkollektiv (1987): Einführung in die Methodik des Feldversuchs
Deutscher Landwirtschaftsverlag, Berlin, 2. Auflage
- COCHRAN, W.G.; COX, G.M. (1957): Experimental Design
Wiley and Sons, New York
- DÖRFEL, H.; BÄTZ, G. (1973): Die Schätzung der Standardabweichung von Mittelwertdifferenzen
in zwei- und dreifaktoriellen Streifenanlagen
Biom. Zeitschrift 15, 417-430
- DÖRFEL, H.; BÄTZ, G. (1980): Mittelwertvergleiche bei signifikanter Wechselwirkung
Archiv Acker- und Pflanzenbau und Bodenkunde 24, 323-328
- FEDERER, W. T. (1975): The misunderstood split plot
Applied Statistics, North-Holland Publ., 9-39
- MUDRA, A. (1958): Statistische Methoden für landwirtschaftliche Versuche
Springerverlag, Berlin
- THOMAS, E. (1966a): Die Anlage mehrfaktorieller Versuche unter besonderer Berücksichtigung
der Mechanisierung des Feldversuchswesens
A.-Thaer-Archiv 10, 665-680
- THOMAS, E. (1966b): Die Anlage mehrfaktorieller Versuche
A.-Thaer-Archiv 10, 757-767

FALLBEISPIELE

Die im Band 20, Heft 4 dieser Zeitschrift von K. ULM et al. publizierten „Strategien zur Auswertung einer Therapiestudie mit der Überlebenszeit als Zielkriterium“ stellten verschiedene Auswertungsansätze am Beispiel einer konkreten Studie zur Diskussion. Der umfangreiche Kommentar von L. Edler und C. Quintero sowie die entsprechenden Antworten der Autoren hierzu sind die Rückkopplung des „Call for comments“, zu dem im Editorial damals aufgerufen wurde.

Biometrie und Informatik in Medizin und Biologie **22** (2), 59 – 66, ISSN 09 34-9235
© Verlag Eugen Ulmer GmbH & Co., Stuttgart; Gustav Fischer Verlag KG, Stuttgart

Kommentar

zu K. Ulm, C. Schmoor, W. Sauerbrei, G. Kemmler, Ü. Aydemir, B. Müller und M. Schumacher: (1989). Strategien zur Auswertung einer Therapiestudie mit der Überlebenszeit als Zielkriterium. Biometrie und Informatik in Medizin und Biologie, 20, 171-205.

L. Edler und C. Quintero

1. Einleitung

Die Autoren K. ULM, C. SCHMOOR, W. SAUERBREI, G. KEMMLER, Ü. AYDEMIR, B. MÜLLER und M. SCHUMACHER liefern mit ihrem Fallbeispiel der Auswertung der Überlebenszeit in einer Therapiestudie zum malignen Gliom beim Erwachsenen einen außerordentlich wichtigen Beitrag zur problemgerechten Anwendung biometrischer Verfahren bei der Auswertung von Therapiestudien. Darüber hinaus liefert das Beispiel eine Reihe von Ansätzen zur Verbesserung oder Neuentwicklung biometrischer Verfahren.

Da nach dem derzeitigen medizinischen Verständnis des Verlaufs und der Therapiemöglichkeiten bei Krebs die Überlebenszeit der Patienten eindeutig im Vordergrund steht, kommen statistischen Methoden zu effizienten und unverzerrten Auswertungen dieser Zielgröße besondere Bedeutung zu. Aus diesem Grund müssen alle diejenigen Faktoren in einer Auswertung berücksichtigt werden, die einen Unterschied der Überlebenszeit aufzeigen. Da im allgemeinen sowohl die Faktoren selbst als auch die Stärke ihres Einflusses auf die Überlebenszeit nicht hinreichend gesichert ist, muß schon bei der Planung einer Therapiestudie auf eine möglichst vollständige Erfassung bedeutsamer Merkmale geachtet werden. Bei der Auswertung sind biometrische Verfahren einzusetzen, die (a) die prognostisch wichtigen Faktoren erkennen und aus der Gesamtheit aller Faktoren herausfiltern und die (b) einen unverzerrten Therapievergleich gewährleisten, indem gerade diese Faktoren berücksichtigt werden. Diese beiden Punkte werden in der Arbeit für das Cox'sche Regressionsmodell ausführlich behandelt und an einer multizentrischen zwei-armigen klinischen Studie exemplifiziert.

Über die Berücksichtigung des Einflusses von Kovariablen auf den Therapievergleich und die Analyse von Therapieeffekten in prognostischen Untergruppen hinaus, spielen die Ergebnisse einer Therapiestudie auch allgemein für die Bestimmung prognostischer Faktoren eine große Rolle, da in einer kontrollierten Studie erhobene Daten meist von einer besseren Qualität sind als solche, die in retrospektiv angelegten Prognosestudien gesammelt werden. Deswegen haben die vorgestellten Methoden auch erheblichen Modellcharakter für die Auswertung von Prognosestudien. Dieser Aspekt soll im folgenden mit berücksichtigt werden. Es ist schon etwas verwunderlich, daß das Cox'sche Regressionsmodell bisher nur selten unter dem Aspekt praktischer Anwendung dargestellt wurde (BYAR, 1982 und 1984; EDLER, 1986), obwohl es schon fast 20 Jahre bekannt ist, und – nicht nur in der Biomedizin – häufig angewandt wird.

Eine Ursache mag darin liegen, daß es in der Regel schwierig ist, ausführliche methodische Beschreibungen in den Originalarbeiten über klinische Resultate unterzubringen. Deswegen ist besonders anzuerkennen, daß die Herausgeber von *BIOMETRIE UND INFORMATIK IN MEDIZIN UND BIOLOGIE* diese Lücke erkannt haben und offensichtlich mit einer Sektion „Fallbeispiele“ zu füllen gedenken. Wünschenswert ist, daß dies auch von den klinischen Partnern als ergänzende Literatur zu den Originalarbeiten aufgegriffen wird. Das Fallbeispiel zeigt sehr deutlich die Fülle weiterer substantieller medizinischer Resultate oder Zusatzinformationen, die in der medizinischen Originalarbeit sonst nur schwer oder gar nicht zu finden sind. Gerade vergleichenden Analysen zeigen die Stärken und Schwächen der Interpretation klinischer Ergebnisse.

Wenn auch bei dem vorliegenden Beispiel der Schwerpunkt auf der biometrischen Methodenwahl liegt, so werden doch eine Reihe methodischer Hinweise und Anweisungen gegeben, die für in der Analyse von Überlebenszeiten weniger geübte Biometriker von unschätzbarem Wert sein dürften.

Im folgenden werden einige Punkte aufgegriffen, die Anlaß für Fragen und Ergänzungen waren. Wir hoffen, damit weitere Ansatzpunkte zur Diskussion von Auswertestrategien bei klinischen Studien zu geben. Es sollte bedacht werden, daß wir durch zusätzliche oder andere Auswertungen eine wesentliche Änderung in den medizinischen Schlußfolgerungen nicht erwarten und daß wir die Studie als hinreichend ausgewertet ansehen. Denn bei der Auswertung klinischer Therapiestudien muß man sich auch darüber im klaren sein, daß „you can not spend a total life in analyzing one data set“ (Cook & WEISBERG, 1986).

2. Planung und Durchführung der Studie.

Bei der Beschreibung des Patientenkollektives fällt auf, daß unter den 447 aus insgesamt 17 Kliniken in die Studie randomisierten Patienten keine Protokollverletzungen oder frühe Ausfälle (z. B. Therapieverweigerung nach Randomisation) auftraten. Warum wurde die Studie vor Erreichen der geplanten $n=500$ Patienten beendet und wie wurde die Randomisation geschichtet nach 17 Kliniken durchgeführt? Dies stellt gemeinhin in der Praxis ein Problem dar, bei dem man zwischen einer Balanzierung innerhalb eines Stratum und einer Gesamtbalanzierung abzuwägen hat, es sei denn man wählt eine Blockrandomisation mit sehr kleiner Blocklänge mit dem Nachteil, daß dies alternierenden Reihen nahe kommt.

In der vorliegenden Studie nahmen 6 Kliniken mit jeweils weniger als 10 Patienten teil! Wie groß war die Unbalance pro Klinik? Dies ist wegen Unterschieden zwischen den Kliniken von Interesse.

3. Verborgene Kovariable

Die Modellspezifikation, d. h. hier die Einbeziehung aller einflußreichen Kovariablen, ist bei der Anwendung von Regressionsmodellen ein besonderes Problem. Da in einer klinischen Therapiestudie stets nur eine begrenzte Anzahl von Kovariablen ausreichend dokumentiert werden kann, und da bei dem begrenzten medizinischen Wissen über Krankheit und indizierte Therapie wichtige – aber noch unbekannte – Faktoren unbeachtet bleiben, ist dies eine fast unlösbare Aufgabe.

Die beobachteten Klinikunterschiede legen die Existenz verborgener Kovariablen auch in dieser Studie nahe. Jedenfalls konnten diese Unterschiede nicht durch eine ungleiche

Verteilung bekannter prognostischer Faktoren erklärt werden. Über die Analyse der Klineffekte mit dem Cox-Modell (Punkt 1. in Kap. 3.1.4) hinaus könnte eine Untersuchung der prognostischen Faktoren auf Heterogenität zwischen den Kliniken weiteren Aufschluß geben. Bereits einfache Kontingenztafeln der Häufigkeiten in den verschiedenen Kliniken reichen dazu aus.

Neben der zitierten Literatur haben auch SOLOMON (1984) und STRUTHERS und KALBFLEISCH (1987) Beiträge zur Fehlklassifikation bei Überlebenszeitmodellen geliefert.

4. Fehlende Werte

Fehlende Werte sind in biomedizinischen Untersuchungen selten vermeidbar und müssen deswegen im Modell und im Auswerteverfahren berücksichtigt werden. Die Zurückhaltung und Kritik der Autoren bezüglich des Ersetzens fehlender Werte durch Schätzverfahren kann nur unterstützt werden. Jedoch müssen deshalb Patienten mit fehlenden Werten in einer Kovariablen nicht schon vornherein aus der Analyse ausgeschlossen werden! Die Tatsache einer fehlenden Angabe kann stattdessen als weitere Kategorie definiert werden, womit diese Fälle in der Auswertung verbleiben könnten (EDLER, 1986). Auf diese Weise bliebe die Kovariable „Tumorrest“ im Modell. Der Nachteil des Ersetzens soll nicht verschwiegen werden: Erhöhung der Anzahl von Dummy-Kovariablen und Beschränkung auf eine nur nominale Skala. Dies ist von besonderem Nutzen bei der univariaten Analyse und der Auswertung des Endmodelles. Die Prognose (d. h. das relative Risiko beurteilt durch den Regressionskoeffizienten) der Kategorie „fehlender Wert“ kann zusätzlich über Art und Effekt des Fehlens Aufschluß geben.

5. Faktorensélection

Der Empfehlung der Autoren, für die Variablenselektion sowohl univariate als auch multiple Regressionen durchzuführen, sollte in der Praxis unbedingt gefolgt werden. Einem Vorschlag von L. Hothorn folgend wollen wir hier die der Regressionssituation besser gerecht werdende Bezeichnung „multiple“ anstelle von „multivariate“ benutzen, da nur **eine** Überlebenszeit ausgewertet wird. Wie das Beispiel in Tabelle 1 zeigt, vermag eine multiple Regression bedeutsame aber korrelierte Faktoren u. U. nur schwer zu identifizieren.

Die geschätzten Standardabweichungen zeigen, wie die beiden systematisch korrelierten Kovariablen Leukozyten und Granulozyten sich die Macht im Wald'schen Test streitig machen. Dies ist ein extremes Beispiel, da Leukozyten und Granulozyten systematisch hoch korreliert sind; aber es zeigt deutlich den Effekt von Korrelationen auf die Selektion.

Insgesamt ist zu erwarten, daß die multiple Regression wenig zur Variablenselektion beitragen wird, da die univariat auffallenden Variablen in der multiplen Analyse meist höhere Standardabweichungen ihrer Regressionskoeffizienten zeigen. Dies trifft auch auf das vorliegende Fallbeispiel (siehe Tabelle 3.1.2) zu. Interessant wäre ein Datensatz, in dem das umgekehrte eintritt: niedrigere Varianz in der multiplen Regression als in der univariaten.

Ein anderes Problem, das in der Arbeit nur am Rande angesprochen wird, ist die Wahl des Signifikanzniveaus α^* , welches eine Variable mit ihren p-Wert zu unterschreiten hat, um in die nachfolgende Analyse aufgenommen zu werden. α^* sollte nach dem eben

Tabelle 1: Inflation der Varianz und Reduktion der Macht bei der Variablenselektion in der multiplen Cox-Regression durch zwei korrelierte Kovariablen bei der Bestimmung prognostischer Faktoren zum fortgeschrittenen kolo-rektalen Karzinom (EDLER et al, 1986).

Modellparameter	Leukozyten	Granulozyten	Geschlecht	Allgemeinzustand	
				AZ=1	AZ>1
(a) Nur Leukozyten (Logarithmus der Anzahl)					
Beta	0.88	-	- 0.69	1.29	0.80
Varianz	0.08	-	0.06	0.12	0.25
p-Wert	0.002	-	0.004	0.005	
(b) Nur Granulozyten (Logarithmus der Anzahl)					
Beta	-	0.60	- 0.72	1.22	0.72
Varianz	-	0.06	0.06	0.12	0.25
p-Wert	-	0.01	0.002	0.001	
(b) Leukozyten und Granulozyten (Logarithmus der Anzahl)					
Beta	1.49	0.49	- 0.80	1.21	0.70
Varianz	0.66	0.37	0.06	0.12	0.25
p-Wert	0.07	0.42	0.001	0.001	

angesprochenen in der univariaten Analyse kleiner gewählt werden als in der multiplen, um die multiple Auswertung mit nicht zu vielen und eventuell multipl korrelierten Faktoren zu starten. Interessant ist bei der Gliomstudie in diesem Zusammenhang der Faktor Aphasie, der univariat weniger auffällt als multipl und der in der multiplen Auswertung noch vor den beiden Kovariablen Anamnesedauer und Bewußtseinsstörungen rangiert. Diese beiden Faktoren wurden in die multiple Analyse aufgenommen, zeigten aber keinen nennenswerten Beitrag zur Erklärung von Unterschieden in der Überlebenszeit. Dies bestätigt unsere Erfahrung, nach welcher für die Ergebnisse der univariaten Auswertung durchaus schärfere Schranken als 0.05 oder 0.01 angelegt werden können. Dies scheint ein Feld für weitere Simulationsstudien zu sein! Entsprechende Überlegungen ergeben sich auch zur Auswahl von Wechselwirkungstermen. Die Frage, ab wann Variablen so stark korreliert anzusehen sind, daß ihre Aufnahme in eine multiple Regression fragwürdig wird, dürfte schwer zu beantworten sein, da klinische Variable wegen inherent hoher Variabilität zwischen Patienten in der Regel nicht besonders hoch korreliert sind. Folglich könnte eine Korrelation von $r=0.35$, wie sie in dem Beispiel maximal auftrat, durchaus relevant sein. Man wird für eine befriedigende Antwort nicht ohne klinisch inhaltliche Betrachtungen auskommen.

Zwei letzte Punkte zur Variablenselektion:

(i) Es kann durchaus sinnvoll sein, auch eine in einer einzelnen Studie weniger signifikante Kovariable im Endmodell zu belassen, wenn es eine sehr genau und ohne fehlende Angaben meßbare Variable ist, oder wenn sie in der medizinischen Literatur als prognostisch bedeutsam diskutiert wird. Dies erleichtert den Vergleich von Studienergebnissen.

(ii) Bei Vorliegen einer großen Anzahl von Kovariablen ist eine Variablenselektion nützlich, bei welcher zunächst innerhalb bestimmter Klassen von Variablen vorselektiert wird (z. B. klinische Variable, Laborwerte, Parameter der Primärerkrankung, Parameter aus der Diagnose der fortgeschrittenen Erkrankung, epidemiologische Faktoren etc.). Dieses Vorgehen wurde von DRZEWIECKI & ANDERSEN (1981) für klinische Variablen und pathologische Variablen gewählt.

6. Wechselwirkungen

Auf die Bedeutung von Wechselwirkungen wird in dem Beitrag zurecht wiederholt hingewiesen und sie werden unseres Wissens hier zum ersten Mal sehr klar abgehandelt. Vermißt haben wir die Behandlung von Wechselwirkungen höherer Ordnung sowie die Möglichkeit, quantitative Variable zusätzlich noch in einer anderen als der auf den Studienbögen dokumentierten Skala in das Modell aufzunehmen. So ist zum Beispiel zu überlegen, ob statt Alter nicht etwa $(\text{Alter})^2$ oder $\ln(\text{Alter})$ alternative Kovariablen sind, die Unterschiede in der Überlebenszeit besser zu beschreiben vermögen als Alter in seiner Urskala. Diese Frage stellt sich außer bei Zeiten insbesondere bei Laborwerten, wenn sie quantitativ ins Modell einbezogen werden.

Bei der Analyse von Wechselwirkungen ist ebenfalls zu bedenken, daß diese multiplikativ definiert werden. Das heißt, wenn Therapie die Ausprägungen 1 und 2 und der Karnofsky-Index die Ausprägungen 1 und 2 hat, ergibt sich eine Variable mit den Ausprägungen 1, 2 und 4. Hat man beide Variablen mit 0 oder 1 kodiert, erhält man nur die Wechselwirkungsausprägungen 0 und 1. Deswegen ist es interessant, wie in einer Auswertung die Wechselwirkungen im einzelnen definiert werden. Wurden qualitative (Therapie-) Variable mit quantitativen (z.B. Alter) einfach multipliziert? Eine anschauliche Beurteilung von Wechselwirkungen an Hand entsprechend differenzierter Überlebenskurven ist für den Biometriker hilfreich; für die Diskussion mit dem Kliniker dürfte sie aber unentbehrlich sein. Im Beispiel wurde dies für die beiden Faktoren Therapie und Karnofsky-Index aufgezeigt.

Liegen – wie bei der Gliomstudie – Wechselwirkungen zwischen der Behandlung und Kovariablen vor, so kann ein stratifiziertes Regressionsmodell angewandt werden. Thall und Lachin (1986) modellieren Stratumeffekte auf vier Arten: (A) als multiplikativen Effekt auf die Grundhazardfunktion, (B) als Wechselwirkung mit der Grundhazardfunktion, (C) als Wechselwirkung mit der Grundhazardfunktion und anderen Kovariablen und (D) als Wechselwirkung mit den Kovariablen bei gleichzeitigem Effekt auf die Grundhazardfunktion. In der Schreibweise von Ulm et al (1989) ist im Modell C die Hazardfunktion für einen Patienten aus Karnofsky-Stratum s mit Kovariablenvektor z_s gleich

$$\lambda(t, Y, z_s) = \lambda_0(t, s) \exp(\beta_{0,s} Y + \beta_s z_s)$$

mit stratum-spezifischen Regressionskoeffizienten. Ohne die Kovariable z ist dies das Modell im Anhang 2 der Arbeit. Modell B erhält man aus Modell C, indem man Stratumeffekte nur auf die Grundhazardfunktion erlaubt und die β_s über die Strata hinweg als gleich voraussetzt. Ein modifiziertes Modell wurde bereits von Schlichting et al (1983) in einer Therapiestudie zur Behandlung von Leberzirrhose mit/ohne Prednison angewandt. Die Regressionskoeffizienten werden dort stratumweise geschätzt und auf Unterschiede getestet. Unterscheiden sie sich zwischen den Strata nicht und sind sie signifikant von Null verschieden, werden sie als prognostische Kovariable ins Modell aufgenommen (β^P). Sind sie signifikant unterschiedlich, werden sie als stratumspezifische („therapeutische“) Variable aufgenommen (β^T). Daraus resultiert ein Cox-Modell der Form

$$\lambda_0(t) \exp \left(\beta_0 Y + \beta_{s1}^T Z_{11} + \dots + \beta_{sk}^T Z_{k1} + \beta_{s(k+1)}^P Z_{k+1} + \dots + \beta_{s(k+r)}^P Z_{k+r} \right)$$

Stratumspezifische Modelle eröffnen eine weitere Dimension für die Step-down Modell-suche im Cox-Modell und wurden von Thall & Lachin (1986) für eine klinische Studie beim Prostatakarzinom benutzt.

Die Analyse der Klinikeffekte in Kapitel 3.1.4 deutet auf die Existenz der eingangs erwähnten verborgenen Kovariablen hin. Warum wurde Klinik – eventuell nach geeigneter Zusammenfassung – nicht als weitere Kovariable in das Endmodell aufgenommen? Durch die Aufsplittung in 11 Klinikstrata besitzt ein Test auf Klinikeffekte und auf Wechselwirkungen zwischen Klinik und Therapie u. U. zu wenig Macht im Vergleich zu den übrigen dichotomisierten Einflußgrößen. Die in Kapitel 4 beschriebene CART-Methode könnte direkt auf die Zusammenfassung der insgesamt 17 Kliniken (nominales Merkmal!) angewandt werden und danach zum Beispiel eine dichotomisierte „Klinik-Variable“ generiert werden.

7. Bootstrap-Resampling

Ein Aspekt des Bootstrap-Verfahrens, der in der Aufgabenstellung der Autoren allerdings von untergeordneter Priorität war, ist die damit mögliche Modelldiagnostik. Bootstrap-Replikationen, bei denen Kovariable im Modell bleiben, die sonst nur in wenigen Bootstrap-Replikationen ins Endmodell aufgenommen werden, entsprechen der Auswahl einer Patientenpopulation, die wenig repräsentativ für die vorliegende Studienpopulation ist. Mit anderen Worten, Bootstrap-Replikationen mit Kovariablen, die selten auftauchen, sind solche, in denen bevorzugt Patienten ausgelassen wurden, die von hohem Einfluß auf das Ergebnis sind.

Diese Überlegung führt zur Frage: Gibt es in dem Kollektiv der ausgewerteten 447 Patienten Fälle die von hohem Einfluß auf den Schätzwert β der Regressionskoeffizienten sind. Wenn ja, welche sind es? Und vor allem, wie ändert sich β und welchen Einfluß hat dies auf die Variablenselektion und das Endmodell, wenn diese Fälle eliminiert werden. Antwort darauf gibt ein einfaches „Leaving-One-Out“ Verfahren (EDLER, 1986) oder das von STORER und CROWLEY (1985) vorgeschlagene Verfahren einer Approximation der Likelihood-Funktion (Ein-Schritt Newton-Raphson). Eine ausführliche Darstellung des Regressionsdiagnostik für das Cox-Modell und die Bestimmung des Einflusses einzelner Fälle findet man auch bei PETTITT & DAUD (1989). Gewisse Schwierigkeiten im Verständnis der von Ulm et al (1989) durchgeführten Bootstrap-Methode bereitet die Vermischung von zwei Zielsetzungen: Kodierung und Selektion. Durch die Einführung von korrelierten Dummy-Variablen wie z. B. Karnofsky-Dummy 1: < 70 , ≥ 70 und Karnofsky-Dummy 2: < 80 , ≥ 80 kann die Varianzschätzung von $\beta_{\text{Karnofsky}}$ erhöht und damit die Aufnahmehäufigkeit des Faktors Karnofsky-Index erniedrigt werden. Dies könnte eine Erklärung dafür sein, daß seine Aufnahmehäufigkeit mit 66% näher bei der Häufigkeit des nicht signifikanten Organischen Psychosyndroms (56%) liegt als bei dem vergleichbar bedeutsamen Malignitätsgrad (88%). Es stellt sich die Frage ob die Kodierung nicht gänzlich von der Variablenselektion getrennt und vorgezogen werden muß. Im Fall der Beurteilung von Therapieeffekten mag die Antwort offen sein; im Fall einer Prognosestudie muß sie jedoch der Verallgemeinerung wegen bejaht werden.

Problematisch erscheint der Versuch, aus den Bootstrap-Ergebnissen mit einer Wiederholungsanzahl von $M = 100$ aus der Aufnahmehäufigkeit auf die Korrelation von Kovariablen schließen zu wollen. Besser erscheint Korrelation und Multikollinearität zunächst an Hand der Kovariablenwerte direkt zu bestimmen, wobei allerdings wegen der hohen Variabilität klinischer Parameter auch zahlenmäßig geringe Korrelationen bedeutsam sein können. Eine Wiederholungsanzahl von $M = 100$ erscheint im Vergleich zu den gegenwärtig bei Bootstrap-Verfahren diskutierten $N = 1000 - 2000$ zu niedrig, wäre mindestens aber zu begründen.

8. CART Methode

Die Anwendung der Klassifikations- und Regressionsbäume weist über die Fragestellung der Autoren insofern hinaus, als es eine Methode zu Bestimmung prognostischer Untergruppen ist (vgl. Abb. 42), deren Festlegung im vorliegenden Fall allerdings von der Therapie abhängt. Insofern wäre es interessant der Unterteilung in Prognosegruppen durch CART einer Unterteilung durch das Cox-Modell gegenüber zu stellen, in dem entweder die gleichen Kovariablen wie bei CART oder die im Cox-Endmodell erhaltenen Kovariablen benutzt werden. Daraus wäre weiterer Aufschluß über die Rolle des Karnofsky-Index zu erwarten (Korrelation mit Alter). Erwarten würde man eine höhere Varianz des β -Schätzers für den Therapieeffekt bei der CART-Klassifikation.

9. Prognoseindex / Therapieeffekt

Die therapeutische Wirkung wird als relatives Risiko $\exp(\beta_0 + \delta z)$ bzw. log-relatives Risiko $\beta_0 + \delta z$ beschrieben, wobei δ der Koeffizientenvektor für die Wechselwirkungen der Kovariablen z mit der Therapie ist. Diese Quantifizierung entspricht der Berechnung eines Prognoseindex $I(z) = \beta z$ in dem unter β und z die Regressionsparameter und Kovariablen des Endmodells sind (ANDERSEN et al, 1983; EDLER et al, 1986). ANDERSEN et al (1983) berechnen zusätzlich den Median als anschauliches Maß für eine Beurteilung von Unterschieden in der Überlebenszeit. Dies wäre im vorliegenden Fall eine gute Ergänzung. Allerdings handelt man sich dabei eine größere Variabilität ein als beim relativen Risiko. Ein Vergleich verschiedener Prognosemodelle ist kürzlich von BONSEL et al (1990) für die Beurteilung der Wirksamkeit von Lebertransplantationen zur Behandlung von Zirrhosen vorgestellt worden.

Auf ein Problem im Rahmen von Prognosestudien zur Validierung der TNM-Klassifikation soll am Schluß noch hingewiesen werden. Dort treten korrelierte Kovariable und fehlende Angaben besonders unangenehm auf. Die klinische TNM-Klassifikation und die post-operative TNM-Klassifikation (p-TNM) sind naturgemäß hoch korreliert. In beiden Klassifikationen ist mit fehlenden Werten zu rechnen. p-TNM ist als die validere Klassifikation anzusehen, liegt aber in der Regel für weniger Patienten vor als die klinische und ihr Fehlen könnte ein Indiz für eine schlechtere Prognose sein. Entsprechende Probleme bei der Konstruktion eines Cox-Modells zur Prognose treten bei Vorliegen von Informationen aus „second-look“ Operationen oder Restagings im Verlauf einer Chemotherapie auf.

Literatur

- ANDERSEN PK, CHRISTENSEN E, FAUERHOLDT L, SCHLICHTING P (1983): Measuring prognosis using the proportional hazards model. *Scand J. Statist.* 10, 49-52.
- BONSEL GJ, KLOMPMAKER IJ, VAN'TVEER F, HABBEMA JDF, SLOOFF MJH (1990): Use of prognostic models for assessment of value of liver transplantation in primary biliary cirrhosis. *The Lancet*, 335, 493-497.
- BYAR DP (1982): Cox and Weibull models with covariates. In: MIKE V, STANLEY KE (eds): *Statistics in Medical Research. Methods and Issues, with application in Cancer research*. New York, Wiley 1982, 365-401.
- BYAR DP (1984): Identification of prognostic factors. In: BUYSE ME, STAQUET MJ, SYLVESTER RJ (eds): *Cancer clinical trials-methods and practice*. Oxford: Oxford University Press, 1984, 403-406.
- COOK RD, WEISBERG S (1982): *Residuals and Influence in Regression*. London, Chapman and Hall.

- DRZEWIECKI KT, ANDERSEN PK (1981): Survival with malignant melanoma. A regression analysis of prognostic factors. *Cancer* 49, 2414-2419.
- EDLER L (1986): Auswertungsstrategien mit dem Cox'schen Regressionsmodell zur Bestimmung prognostischer Faktoren. In: EHLERS, CT, BELAND H (eds). *Perspektiven der Informationsverarbeitung in der Medizin*. 31. Jahrestagung der GMDS, Göttingen, Sept. 1986. Springer, Berlin, 403-406.
- EDLER L, HEIM ME, QUINTERO C, BRUMMER T, QUEISSER W (1986): Prognostic factors of advanced colorectal cancer patients. *Eur. J. Cancer. Clin. Oncol*, 22, 1231-1237.
- PETTIT AN, DAUD IB (1989): Case-weighted measures of influence for proportional hazards regression. *Appl. Statist.* 38, 51-67.
- SCHLICHTING P, CHRISTENSEN E, ANDERSEN PK, FAUERHOLDT L, JUHL E, POULSEN H, TYGSTRUP N, and the Copenhagen Study Group for liver diseases (1983): Prognostic factors in Cirrhosis identified by Cox's regression model. *Hepatology*, 3, 889-995.
- STORER BE, CROWLEY J (1985): A diagnostic for Cox regression and general conditional likelihoods. *J. Amer. Statist. Assoc.* 80, 139-47.
- STRUTHERS CA, KALBFLEISCH JD (1986): Misspecified proportional hazard models. *Biometrika*, 73, 363-9.
- SOLOMON PJ (1984): Effect of misspecification of regression models in the analysis of survival data. *Biometrika* 71, 431-434.
- THALL PF, LACHIN JM (1986): Assessment of stratum-covariate interactions in Cox's proportional hazards regression model. *Statistics in Medicine*, 5, 73-83.
- ULM K, SCHMOOR C, SAUERBREI W, KEMMLER G, AYDEMİR Ü, MÜLLER B, SCHUMACHER M (1989): Strategien zur Auswertung einer Therapiestudie mit der Überlebenszeit als Zielkriterium. *Biometrie und Informatik in Medizin und Biologie* 20, 171-205.

Anschrift der Verfasser: Dr. Lutz Edler, National Institute of Environmental Health Sciences (NIEHS), P.O. Box 12233, MD B 302, Research Triangle Park, NC 27709, USA · Dr. Constanza Quintero, Departamento de Matemáticas y Estadística Universidad, Nacional de Colombia, Bogotá, Kolumbien

Antwort auf den Kommentar von L. Edler und C. Quintero

zu K. Ulm, C. Schmoor, W. Sauerbrei, G. Kemmler, Ü. Aydemir, B. Müller und M. Schumacher (1989): Strategien zur Auswertung einer Therapiestudie mit der Überlebenszeit als Zielkriterium, Biometrie und Informatik in Medizin und Biologie 20, 171-205.

K. Ulm

1. Planung und Durchführung der Studie

Die Daten der 447 Patienten, die unserer Publikation zugrunde liegen, beziehen sich auf eine Zwischenauswertung mit dem Stichtag 1.1.88. In der Zwischenzeit wurden die geplanten 500 Patienten in die Studie aufgenommen. Die Ergebnisse geben also nicht die Resultate der Endauswertung wieder. Der Abschlußbericht liegt in der Zwischenzeit vor. Es gab einige wenige Ausschlüsse von Patienten, bei denen nachträglich ein malignes Gliom vom Grad 2 diagnostiziert wurde. Einige Patienten ($n = 18$) brachen die Therapie wegen schwerer Nebenwirkungen ab. Diese Ausschlüsse und Abbrüche waren in beiden Therapiegruppen in etwa gleich verteilt.

2. Faktorensélection

Auf das Problem von korrelierten Variablen bei einer multiplen Regression (Multikollinearität) wurde von vielen Autoren auch in anderen Zusammenhängen bereits hingewiesen (z. B. KLEINBAUM et al., 1982). So ist es keine Überraschung, daß in dem Beispiel von EDLER und QUINTERO die Leukozyten den Vorzug vor den Granulozyten erhalten. Bei der eindimensionalen Analyse ist deren p-Wert entsprechend niedriger. In einem schrittweisen Vorgehen würde nur die Variable „Leukozyten“ Berücksichtigung finden. Das Problem der Variablenselektion wurde z. B. auch im Zusammenhang mit der Diskriminanzanalyse intensiv behandelt. Von BOCK, POLLOK, SCHACH et al. (1976) wurde z. B. auch die von EDLER und QUINTERO beschriebene Situation der geringeren Varianz bei der multiplen Regression erwähnt, also der Fall, in dem zwei Variable, jede für sich betrachtet, keinen Einfluß auf das Zielmerkmal ausgeübt, gemeinsam jedoch ein Zusammenhang besteht.

Diese Multikollinearität ist natürlich auch bei der CART-Analyse zu berücksichtigen. Denn hier werden ja die beiden gesplitteten Untergruppen getrennt weiter analysiert. So ist es durchaus denkbar, daß bei einer leicht geänderten Datenlage eine zweite korrelierte Variable zuerst ausgewählt wird und die weitere Analyse zu einem komplett anderen Ergebnis führt.

3. CART-Methode

Die CART-Methode ist, wie auch EDLER und QUINTERO bemerken, primär zur Bestimmung prognostisch homogener Untergruppen gedacht. Neu in unserer Publikation war die Anwendung zur Beurteilung von Therapieeffekten. Auf den Vergleich zwischen der CART-Methode und dem COX-Modell sind SEGAL und BLOCH (1989) eingegangen. Das Ergebnis beider Ansätze ist nicht direkt miteinander vergleichbar. SEGAL und BLOCH (1989) beobachten ebenfalls das gleiche Phänomen, daß sich Merkmale im Cox-Modell als signifikant erweisen, in der CART-Analyse jedoch nicht in Erscheinung treten und umgekehrt. Ein wesentlicher Grund liegt in den Wechselwirkungen, die modellmäßig nur schwierig in der Form des CART-Ansatzes zu beschreiben sind, vor allem ohne Kenntnis dieses Ergebnisses. Vielmehr kann das Ergebnis des CART-Ansatzes dazu verwendet werden, das COX-Modell näher zu spezifizieren. Da sich der CART-Ansatz sehr an den vorhergehenden Daten orientiert, ist es dringend erforderlich, das Ergebnis an neuen Daten zu verifizieren.

Der Karnofsky-Index tritt beim CART-Ansatz nicht als Split-Merkmal in Erscheinung. In der Gruppe der unter 50jährigen wird zuerst anhand des Malignitätsgrades weiter aufgeteilt. Hier erweist sich der Karnofsky-Index als nächst besseres Split-Merkmal. Der Log-Rank-Test erreicht einen Wert von 9.5 anstelle von 17.72 beim Malignitätsgrad. Das Ergebnis der CART-Analyse wird auch durch die Abbruchkriterien beeinflusst. Zwei Kriterien wurden verwendet: a) Ergebnis der Log-Rank-Tests und b) Fallzahl pro Untergruppe. Eine Veränderung dieser beiden Kriterien kann zu unterschiedlichen Regressionsbäumen führen. Ein möglicher Ansatz wäre daher, die CART-Analyse mit verschiedenen Abbruchkriterien durchzuführen.

Literatur

- BOCK, P.R.; POLLOCK, B.; SCHACH, S.; FUCHS, A.; LOHAUS, R.: Classification of Psychoactive Drugs by Visually Evoked Potentials in Rabbits by Means of Multiple Discriminant Analysis. *Arzneim.-Forsch. (Drug Res.)* 26, Nr. 7 (1976), 1308-1320
- KLEINBAUM, D.G.; KUPPER, L.L.; MORGENSTERN, H.: *Epidemiologic Research*. Van Nostrand Reinhold Company New York, 1982
- SEGAL, MR; BLOCH, DA: A comparison of estimated proportional hazard models and regression trees. *Statistics in Medicine* 8, 1989, 539-550.

Antwort auf den Kommentar von L. Edler und C. Quintero

zu K. Ulm, C. Schmoor, W. Sauerbrei, G. Kemmler, Ü. Aydemir, B. Müller und M. Schumacher: (1989). Strategien zur Auswertung einer Therapiestudie mit der Überlebenszeit als Zielkriterium. Biometrie und Informatik in Medizin und Biologie, 20, 171-205.

C. Schmoor

1. Analyse der Klinikeffekte

Wie in Abbildung 2.3. der Arbeit von ULM et al. (1989) dargestellt, zeigen sich zwischen den einzelnen Kliniken starke prognostische Unterschiede. Mögliche Ursachen dafür sind

- eine unterschiedliche Zusammensetzung der Patientenkollektive der einzelnen Kliniken bzgl. bekannter prognostischer Faktoren,
- eine unterschiedliche Zusammensetzung bzgl. unbekannter prognostischer Faktoren („verborgene Kovariable“),
- ein Einfluß der Kliniken selbst auf die Prognose der Patienten.

Die beiden letztgenannten Ursachen für die beobachteten Klinikeffekte werden auch durch sorgfältigste Analysen nicht zu unterscheiden sein, während die Zusammensetzung der Patientenkollektive bzgl. bekannter prognostischer Faktoren und deren Einfluß auf prognostische Unterschiede untersucht werden kann. Imbalancen zwischen den Kliniken bestehen bzgl. der prognostisch bedeutsamen Variablen Alter, Karnofsky-Index, Malignitätsgrad, Resektion und Cortison. Wie sich bei den Analysen gezeigt hat, sind jedoch auch dann noch prognostische Unterschiede zwischen den Kliniken nachzuweisen, wenn diese Imbalancen berücksichtigt werden, d.h. diese prognostischen Faktoren ins Cox-Modell miteinbezogen werden. Das deutet darauf hin, daß die Imbalancen bzgl. der bekannten prognostischen Faktoren nicht die einzige Ursache für die prognostischen Unterschiede zwischen den Kliniken ist.

Grundsätzlich besteht natürlich in einer Multicenter-Studie, an der viele Kliniken teilnehmen, die teilweise nur wenige Patienten einbringen, bei der Analyse von Klinikeffekten das Problem, diese überhaupt sinnvoll zu modellieren. Bei unseren Analysen wurde für jede der 11 größeren Kliniken mit mehr als 10 Patienten der Effekt einzeln modelliert und die übrigen 6 Kliniken zu einer Gruppe zusammengefaßt. Da eine inhaltlich begründete Zusammenfassung der Kliniken nicht vorliegt, scheint dieses eine sinnvolle Modellierung zu sein, obwohl ein simultaner Test von 11 Dummy-Variablen im Vergleich zum Test einer dichotomen Einflußgröße natürlich eine geringere Power aufweist. Von EDLER und QUINTERO wird nun vorgeschlagen, etwa mit Hilfe der Cart-Methode

die 17 Kliniken in zwei oder auch mehrere prognostisch in sich homogene Gruppen zusammenzufassen und diese so neu gebildete Variable in die Analyse miteinzubeziehen. Dieses könnte einerseits der Adjustierung der Therapieeffekte und der Effekte der anderen prognostischen Faktoren bzgl. der Kliniken und andererseits der Beurteilung der prognostischen Unterschiede zwischen den Kliniken oder deren Wechselwirkung mit der Therapie dienen. Die Bildung und Einbeziehung solch einer neuen „Klinikvariable“ scheint jedoch keine neuen Erkenntnisse zu liefern. Eine Adjustierung der anderen Effekte wurde bereits wie in Abschnitt 3.1.4. beschrieben durch die zusätzliche Aufnahme der oben genannten 11 Dummy-Variablen ins Endmodell durchgeführt. Dabei zeigten sich für die anderen Effekte – im Vergleich zu der Analyse, die die Kliniken nicht berücksichtigt, – keine qualitativ unterschiedlichen Ergebnisse. Eine Beurteilung der prognostischen Unterschiede zwischen den Kliniken mit Hilfe dieser neu gebildeten „Klinikvariable“ ist einerseits problematisch, da sie bereits datenabhängig basierend auf dem Zielkriterium generiert wurde. Andererseits ist sowohl die prognostische Bedeutung als auch die Wechselwirkung dieser Variablen mit der Therapie inhaltlich nicht sinnvoll zu interpretieren.

2. Faktorensélection

Bei der Auswertung einer Therapiestudie dient die Analyse der prognostischen Faktoren in erster Linie der Selektion der Faktoren, bzgl. derer der Therapieeffekt im Endmodell zu adjustieren ist. Die Evaluierung der prognostischen Bedeutsamkeit der Faktoren selbst, wie es Aufgabe einer sogenannten Prognosestudie ist, ist dabei erst in zweiter Linie von Interesse.

Wir vertreten deshalb bei der Selektion der prognostischen Variablen das Prinzip, lieber eine unbedeutende Variable zu viel zu berücksichtigen als eine bedeutsame Variable zu übersehen. Der Grund dafür ist, daß die Nicht-Berücksichtigung einer prognostisch einflußreichen Variable zu einer verzerrten Schätzung der anderen Effekte führen kann, da die im Modell implizit getroffene Annahme proportionaler Hazardfunktionen in diesem Falle nicht mehr erfüllt ist (SCHUMACHER et al, 1987). Deshalb sollen die prognostischen Faktoren berücksichtigt werden, die entweder in der univariaten Analyse, also bei alleiniger Betrachtung, oder in der multiplen Regression (diese Bezeichnung trifft den Sachverhalt in der Tat genauer als der Begriff „multivariate Analyse“), also bei gleichzeitiger Berücksichtigung der übrigen Faktoren, einen Einfluß gezeigt haben.

EDLER und QUINTERO bemerken, daß die Schätzer der Faktoren in der multiplen Analyse meist höhere Varianzen haben, und somit zu erwarten ist, daß die multiple Regression keine prognostisch bedeutsamen Variablen identifizieren wird, die nicht bereits univariat auffällig sind. Es ist bei diesen Überlegungen jedoch nicht nur die Varianz sondern auch der geschätzte Effekt selbst von Interesse, da dieser in der univariaten Analyse verzerrt sein kann. In der Gliom-Studie liefert die multiple Regression keine zusätzlichen Variablen; grundsätzlich ist jedoch denkbar, daß der Effekt eines Faktors durch einen anderen prognostisch bedeutsamen Faktor überlagert und deshalb erst in der multiplen Analyse erkannt wird.

Auf der anderen Seite kann die zusätzliche Berücksichtigung unbedeutender Variablen schlimmstenfalls eine Varianzerhöhung der Effektschätzer der anderen prognostischen Faktoren verursachen, die jedoch nur dann auftreten kann, wenn diese Faktoren mit der zusätzlich im Modell enthaltenen unbedeutenden Variable korreliert sind. Das bedeutet, daß die Schätzung des Therapieeffekts, die im Vordergrund der Analyse steht, in einer randomisierten Studie davon nicht betroffen ist.

Aufgrund dieser Überlegungen wurde auch für die univariate Analyse kein kleineres Signifikanzniveau als für die multiple Analyse gewählt, wie es von EDLER und QUINTERO vorgeschlagen wurde. Dadurch wird zwar in Kauf genommen, daß auch unbedeutendere

Variablen wie in der Gliom-Studie Anamnesedauer, Organisches Psychosyndrom und Bewußtseinsstörung im Endmodell enthalten sind; jedoch beeinflusst dies – wie oben erläutert und wie auch durch Herausnahme dieser drei Variablen aus dem Endmodell überprüft – die Schätzung der anderen Effekte nur unwesentlich.

EDLER und QUINTERO sprechen das Problem an, korrelierte Variable gleichzeitig in einem Regressionsmodell zu betrachten. Sie demonstrieren in ihrem Beispiel zur Bestimmung prognostischer Faktoren beim kolo-rektalen Karzinom anhand der hoch-korrelierten Variablen Leukozyten und Granulozyten eindrucksvoll den Anstieg der Varianzen der Parameterschätzer. Solche extremen Auswirkungen wie in diesem Beispiel treten natürlich nur bei sehr hoch korrelierten Variablen auf (ANTES, 1990). EDLER und QUINTERO bemerken, daß eine Korrelation von 0.35 – wie sie in der Gliom-Studie maximal zu beobachten war – für klinische Variablen vergleichsweise hoch ist. Dieser Aspekt ist zwar bei der inhaltlichen Einschätzung der Korrelationen selbst von Bedeutung, für die formale Modellbildung hinsichtlich der Auswirkungen korrelierter Variablen jedoch nicht relevant.

Es wird von EDLER und QUINTERO vorgeschlagen, bei einer großen Anzahl potentiell prognostischer Faktoren zunächst innerhalb sachlich zusammengehöriger Klassen von Variablen vorzuselektieren. Dieses Vorgehen scheint jedoch problematisch zu sein. Auf der einen Seite sind die Variablen innerhalb einer Klasse – etwa der Laborwerte – eventuell hoch korreliert, was zu erhöhten Varianzen der Schätzer führen kann. Auf der anderen Seite werden innerhalb der Klassen die Effekte der Variablen nicht bzgl. eventuell bedeutender prognostischer Faktoren anderer Klassen adjustiert, was zu verzerrten Schätzern führen kann.

3. Wechselwirkungen

EDLER und QUINTERO sprechen die Möglichkeit an, quantitative Variable – wie in der Gliom-Studie das Alter – in transformierter Form ins Modell einzubeziehen. Dies hat sowohl für die Analyse der prognostischen Bedeutung der Variablen als auch für die Analyse der Wechselwirkung der Variablen mit der Therapie Konsequenzen. In Abschnitt 3.1.1. der Arbeit von ULM et al. (1989) wurde darauf hingewiesen, daß bei Berücksichtigung einer stetigen Variable, etwa Z , in ihrer ursprünglichen Form die implizite Annahme getroffen wird, daß Z einen loglinearen Einfluß auf die Überlebenszeit hat, und diese Annahme zu Beginn der Analyse zu überprüfen ist. Zu dieser Überprüfung stehen etwa die Vorschläge von THOMSEN (1988) und von TIBSHIRANI und HASTIE (1987) zur Verfügung, die eine Schätzung der Form des log-relativen Risikos in Abhängigkeit von Z beinhalten.

Von THOMSEN wird vorgeschlagen, das log-relative Risiko in Abhängigkeit von Z zu schätzen, indem die Patienten bzgl. ihrer Werte von Z in $n+1$ Gruppen aufgeteilt werden, und ein Cox-Modell angepaßt wird, das ebensoviele Kovariablen enthält. Dabei mißt der Effekt der i -ten dieser Kovariablen, $i=1,...,n$, die Änderung der Steigung des log-relativen Risikos am i -ten Unterteilungspunkt von Z . TIBSHIRANI und HASTIE schlagen eine Kernschätzung des log-relativen Risikos vor. Dabei wird ein Rechteckskern verwendet, so daß für jeden Wert von Z mit Hilfe eines Local-Likelihood-Verfahrens die Steigung des log-relativen Risikos geschätzt wird. Aus der Anwendung dieser Verfahren lassen sich unter Umständen auch geeignete Transformationen für die Variable ableiten. Beide Ansätze wurden in der Gliom-Studie für die Variable Alter verwendet. Für das Verfahren von THOMSEN wurden die Patienten in 7 Altersgruppen mit den Unterteilungspunkten 35, 45, 50, 55, 60, 65 aufgeteilt. Für das Verfahren von TIBSHIRANI und HASTIE wurde eine Bandbreite von 15 gewählt, d.h. die Schätzung für jedes Alter beinhaltet die 7 jeweils niedrigeren und höheren Altersjahre. Die Ergebnisse beider Verfah-

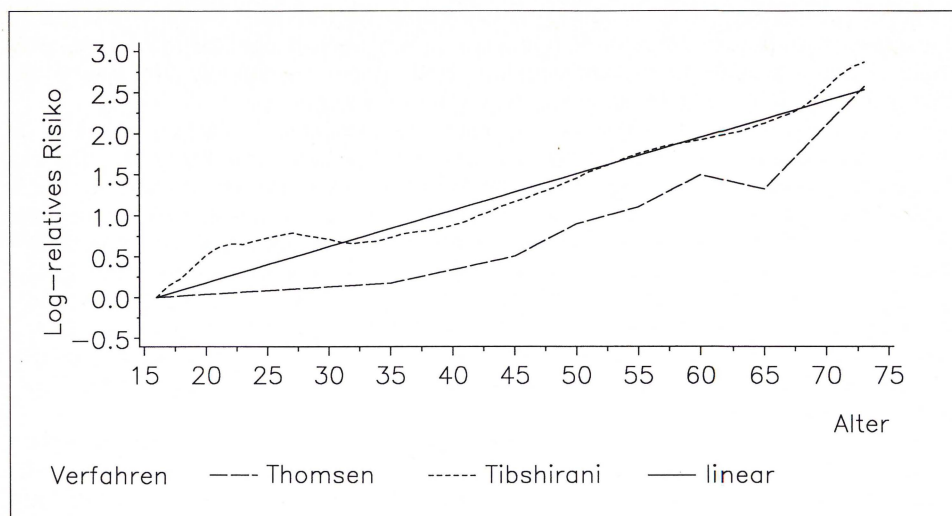


Abbildung 1: Untersuchung der Art des Alterseffekts auf die Überlebenszeit mit Hilfe des Verfahrens von THOMSEN (1988) und TIBSHIRANI und HASTIE (1987)

ren sind in Abbildung 1 dargestellt. Im Vergleich zu der geschätzten Gerade, die man bei Analyse des Alters in seiner ursprünglichen Form im Cox-Modell erhält, zeigen sich keine deutlichen Abweichungen von der Linearität.

Diese Überprüfungen beziehen sich natürlich nur auf die Art des Einflusses einer Variablen auf die Prognose der Patienten. Bei der Modellierung einer multiplikativen Wechselwirkung mit der Therapie wird implizit die zusätzliche Annahme getroffen, daß auch die Interaktion log-linear ist. Diese Annahme könnte überprüft werden, indem man eines der oben angesprochenen Verfahren separat für die beiden Therapiegruppen durchführt. Dieses liefert für jede Therapiegruppe in Abhängigkeit des Alters einen Schätzer für das log-relative Risiko. Die Differenz dieser beiden geschätzten log-relativen Risiken spiegelt gerade die Form der Interaktion wider. Zeigt diese in Abhängigkeit des Alters eine deutliche Abweichung von der Linearität, deutet dies darauf hin, daß die Form der Wechselwirkung durch den Produktterm $\text{Alter} * \text{Therapie}$ nicht adäquat beschrieben wird.

In der in Abschnitt 3.1. beschriebenen Auswertung wurde sowohl bei der Analyse der prognostischen Bedeutung des Alters, als auch bei der Analyse der Wechselwirkung zwischen Alter und Therapie, das Alter in seiner ursprünglichen Form berücksichtigt. Zusätzlich wurde jedoch, um die Stabilität der so erhaltenen Ergebnisse zu untersuchen, eine Wechselwirkungsanalyse durchgeführt, in der das Alter in kategorisierter Form (≤ 50 , $51-60$, > 60) untersucht wurde. Diese Analyse zeigte qualitativ das gleiche Resultat, d.h. auch hier zeigt sich keine starke Wechselwirkung mit der Therapie.

EDLER und QUINTERO sprechen darüberhinaus das Problem der Kodierung qualitativer Variablen bei der Analyse von Wechselwirkungen an. Im Cox-Modell sind jedoch – wie in Abschnitt 3.1.1. und in Anhang 1 dargestellt – nicht die Ausprägungen einer Variablen selbst, sondern nur die Differenzen zwischen den Ausprägungen der Variable von Bedeutung, da immer nur die relativen Risiken von Interesse sind. Dies gilt auch, wenn die Wechselwirkung zwischen zwei Variablen, etwa Therapie Y und Karnofsky-Index Z_1 , mit Hilfe des Produktterms $Y * Z_1$ analysiert werden soll. Unabhängig davon, ob die Ausprägungen der Variablen Y und Z_1 mit 0 und 1 oder mit 1 und 2 kodiert sind, ergibt

sich das relative Risiko zwischen den Therapien – mit der Notation wie in Anhang 1 – als $\exp(\beta_0 + \delta_1 Z_1)$. Bei der Kalkulation der relativen Risiken zwischen den Therapien in den einzelnen durch Z_1 gegebenen Untergruppen muß demnach natürlich berücksichtigt werden, welche Ausprägungen Z_1 annimmt. Es ist jedoch unabhängig davon, wie Y und Z_1 kodiert sind, δ_1 immer als der Wechselwirkungseffekt zu interpretieren, und das Analyseergebnis hängt nicht von der Wahl der Kodierung der Variablen ab. Zu beachten ist jedoch die Modellierung einer Wechselwirkung mit einer Kovariablen, etwa Z_2 , die in mehr als zwei Stufen vorliegt. Hier sind die Differenzen zwischen den einzelnen Ausprägungen von Z_2 von Interesse. Die Ausführungen in Abschnitt 3.1.1. zur Kodierung mehrstufiger ordinaler Variablen sind bei der Analyse von Wechselwirkungen genauso relevant wie bei der Analyse der prognostischen Bedeutsamkeit und lassen sich unmittelbar übertragen. So wird etwa bei der Modellierung einer Wechselwirkung zwischen der Therapie Y und einer dreistufigen Kovariable Z_2 , die mit den Ausprägungen 0, 1 und 2 kodiert ist, implizit vorausgesetzt, daß der Unterschied des Therapieeffekts zwischen Gruppe $Z_2 = 0$ und Gruppe $Z_2 = 1$ von der gleichen Größenordnung ist, wie der Unterschied zwischen den Gruppen $Z_2 = 1$ und $Z_2 = 2$. Eine Situation beispielsweise, in der in den Gruppen 0 und 1 kein Therapieeffekt vorliegt, in Gruppe 2 jedoch die eine Therapie der anderen überlegen ist, wird durch diese Art der Modellierung nicht adäquat beschrieben und kann demzufolge in der Analyse nicht erkannt werden. Dazu müßten sowohl für den prognostischen Faktor als auch für den Wechselwirkungseffekt jeweils zwei Dummy-Variable ins Modell aufgenommen oder separate Schätzungen des Therapieeffekts in den durch Z_2 gegebenen Untergruppen durchgeführt werden. Bei der Analyse der Wechselwirkung zwischen Karnofsky-Index und Therapie in Abschnitt 3.1. der Arbeit von ULM et al. wurde der Karnofsky-Index durch eine Variable mit den Ausprägungen 0, 1 und 2 kodiert, d.h. die beschriebenen Annahmen wurden auch in dieser Analyse getroffen. Eine Überprüfung dieser Annahme wurde jedoch durchgeführt, indem – wie in Abschnitt 3.1.3. beschrieben – in jeder der drei durch den Karnofsky-Index definierten Untergruppen der Therapieeffekt separat geschätzt wurde. Dabei zeigte sich (Tabelle 3.1.5), daß der Wechselwirkungseffekt durch die gewählte Modellierung sehr gut beschrieben wird.

Die von EDLER und QUINTERO angesprochene stratifizierte Regressionsanalyse (THALL und LACHIN, 1986) wurde auch für die Gliom-Studie durchgeführt. Dabei wurden die beiden Therapiegruppen als die Strata aufgefaßt und Wechselwirkungen mit den prognostischen Faktoren mit Hilfe von sogenannten „therapiespezifischen“ Kovariablen $(1-Y) * Z_j$ und $Y * Z_j$ definiert (Modell (D) bei EDLER und QUINTERO und bei THALL und LACHIN, 1986). Zur Analyse aller einfachen Wechselwirkungen zwischen der Therapie Y und den prognostisch bedeutsamen Kovariablen Z_1, \dots, Z_k erhält man mit diesem Ansatz ein Cox-Modell der Form

$$\lambda(t|Y, Z_1, \dots, Z_k) = \lambda_0(t) \exp(\beta_0 Y + (1-Y) * \sum_{j=1}^k \beta_{0j} Z_j + Y * \sum_{j=1}^k \beta_{1j} Z_j)$$

wobei	0	Therapie BCNU,
	$Y =$	
	1	Therapie BCNU + VM26
	Z_j	prognostische Faktoren; $j = 1, \dots, k$;
	β_0	Regressionskoeffizient der Behandlung,
	$\beta_{01}, \dots, \beta_{0k}$	Regressionskoeffizienten der prognostischen Faktoren unter Therapie BCNU,
	$\beta_{11}, \dots, \beta_{1k}$	Regressionskoeffizienten der prognostischen Faktoren unter Therapie BCNU + VM 26.

Damit ergibt sich das relative Risiko (RR) zwischen den Behandlungen als

$$RR(Y=1:Y=0) = \exp(\beta_0 + (\beta_{11}-\beta_{01})Z_1 + \dots + (\beta_{1k}-\beta_{0k})Z_k).$$

Das bedeutet also, daß die Wechselwirkungseffekte δ_j aus Anhang 1 der Arbeit von ULM et al. in diesem Modell den Differenzen $\beta_{1j}-\beta_{0j}$, $j = 1, \dots, k$, entsprechen. Die Analyse eines Wechselwirkungseffekts zwischen der prognostischen Variable Z_j und der Therapie Y kann also in diesem Modell durch einen Test der Hypothese „ $\beta_{1j} = \beta_{0j}$ “ durchgeführt werden.

THALL und LACHIN schlagen vor, als Ausgangspunkt das Modell zu betrachten, in dem alle zu untersuchenden Wechselwirkungen zwischen den prognostisch bedeutsamen Variablen und der Therapie enthalten sind. Zur Eliminierung unwichtiger Effekte wird die folgende iterative Step-Down-Prozedur vorgeschlagen.

Schritt 1: Es werden Tests der Hypothesen „ $\beta_{1j} = \beta_{0j}$ “, $j = 1, \dots, k$, durchgeführt und für Kovariablen mit $|\hat{\beta}_{1j}-\hat{\beta}_{0j}| / \sqrt{\text{Var}(\hat{\beta}_{1j}-\hat{\beta}_{0j})} \leq c_1$ die Parameter β_{1j} und β_{0j} zu einem gemeinsamen Parameter β_j zusammengefaßt.

Schritt 2: In dem durch Schritt 1 reduzierten Modell werden alle therapiespezifischen und zusammengefaßten Effekte der Kovariablen eliminiert, für die $|\hat{\beta}_j| / \sqrt{\text{Var}(\hat{\beta}_j)} \leq c_2$ bzw. $|\hat{\beta}_{ij}| / \sqrt{\text{Var}(\hat{\beta}_{ij})} \leq c_2$ ist, $i = 0, 1, j = 1, \dots, k$.

Schritt 1 und Schritt 2 werden solange wiederholt bis keine Effekte aufgrund der vorgegebenen Signifikanzgrenzen mehr zu eliminieren bzw. zusammenzufassen sind.

THALL und LACHIN wollen in Schritt 1 nur dann die Therapiegruppen bzgl. der Kovariableneffekte zusammenlegen, wenn sehr starke Hinweise auf homogene Therapieeffekte bestehen, um möglichst detaillierte Aussagen über optimale Therapien für verschiedene Patientengruppen treffen zu können. Sie sind also in erster Linie nicht daran interessiert, den Fehler 1. Art – also fälschlicherweise auf heterogene Therapieeffekte zu schließen –, sondern den Fehler 2. Art zu kontrollieren. Deshalb wird vorgeschlagen, für die Tests in Schritt 1 den kritischen Wert $c_1 = 0.5$ zu wählen, was einem Niveau von 0.62 entspricht. Falls die wahre Beziehung $|\beta_{1j}-\beta_{0j}| / \sqrt{\text{Var}(\beta_{1j}-\beta_{0j})} = 2$ lautet, liefert diese Wahl einen Fehler 2. Art von 0.06. Für die Eliminierung von Faktoren im Schritt 2 wird ein Signifikanzniveau von $\alpha = 0.1$ vorgeschlagen.

Bei Durchführung dieser Step-Down-Prozedur für die Gliom-Studie ergibt sich ein Endmodell, dessen Resultate in Tabelle 1 dargestellt sind.

Therapiespezifische Kovariablen sind für die Faktoren Karnofsky-Index und Alter im Modell verblieben, wobei der Effekt des Karnofsky-Indexes in der Therapiegruppe BCNU in Schritt 2 des beschriebenen schrittweisen Verfahrens aus dem Modell eliminiert, also gleich Null gesetzt worden ist. Das bedeutet also, daß im Vergleich zur Untersuchung der Interaktionen in Abschnitt 3.1.1. nach Durchführung dieser explorativen

Tabelle 1: Endmodell nach Wechselwirkungsanalyse (THALL und LACHIN, 1986)

Variable	Koeff.	S.E.
Alter (BCNU)	0.025	0.008
Alter (BCNU + VM 26)	0.044	0.009
Karnofsky (BCNU + VM 26)	-0.567	0.128
Malignitätsgrad	0.461	0.089
Resektion		
partiell	-0.974	0.205
total	-0.153	0.133
Epileptische Anfälle	-0.308	0.142
Cortison	0.263	0.133
Therapie	-0.600	0.671

Tabelle 2: Aufgrund der Wechselwirkungsanalyse (THALL und LACHIN, 1986) geschätzte „optimale“ Therapie

	Geschätzte „optimale“ Therapie (Anteil Patienten in %)		Anzahl Patienten
	BCNU	BCNU + VM 26	
alle Patienten	41	59	422
Karnofsky			
> 80	0	100	104
70-80	28	72	198
< 70	98	2	120
Alter			
≤ 50	23	77	164
51-60	32	68	149
> 60	82	18	109

Analyse zusätzlich ein heterogener Therapieeffekt hinsichtlich des Alters zu berücksichtigen ist.

Es können nun aufgrund dieses Endmodells für jeden Patienten prognostische Indizes unter Therapie BCNU und BCNU + VM 26 berechnet und damit entschieden werden, von welcher der beiden Therapien der jeweilige Patient mehr zu profitieren scheint. Tabelle 2 enthält jeweils den prozentualen Anteil Patienten, für den BCNU bzw. BCNU + VM 26 die geschätzte „optimale“ Therapie ist.

Im Gesamtkollektiv ist für etwas über die Hälfte der Patienten BCNU + VM 26 überlegen, was mit dem Ergebnis, daß global kein signifikanter Therapieunterschied vorliegt, übereinstimmt. Bei Aufspaltung des Kollektivs in die drei durch den Karnofsky-Index bestimmten Untergruppen zeigt sich ebenfalls ein zu Abschnitt 3.1.1. konsistentes Ergebnis. Für Patienten mit einem Karnofsky-Index > 80 ist BCNU + VM 26, für Patienten mit einem Karnofsky-Index < 70 ist BCNU die Therapie der Wahl, während für Patienten in der mittleren Gruppe das Ergebnis indifferent ist. Die Variable Alter ist in stetiger Form in die Analyse eingegangen. Bei nachträglicher Aufteilung des Gesamtkollektivs in die Gruppen Alter ≤ 50, 50–60 und > 60 zeigt sich, daß jüngere Patienten eher von der Kombinationschemotherapie profitieren, während sie älteren Patienten eher schadet. Dieses deutet ähnlich wie bei der Variable Karnofsky-Index jedoch in schwächerer Form auf eine qualitative Wechselwirkung der Therapie mit der Variable Alter hin. Interessant ist nun zu untersuchen, inwiefern heterogene Therapieeffekte bzgl. des Alters bei festem Karnofsky-Index der Patienten bestehen. Dies ist in Tabelle 3 dargestellt.

Tabelle 3: Aufgrund der Wechselwirkungsanalyse (THALL und LACHIN, 1986) geschätzte „optimale“ Therapie

Karnofsky	Alter	Geschätzte „optimale“ Therapie (Anteil Patienten in %)		Anzahl Patienten
		BCNU	BCNU + VM 26	
> 80	≤ 50	0	100	54
> 80	51-60	0	100	30
> 80	> 60	0	100	20
70-80	≤ 50	0	100	71
70-80	51-60	0	100	72
70-80	> 60	100	0	55
< 70	≤ 50	95	5	39
< 70	51-60	100	0	47
< 70	> 60	100	0	34

Es zeigt sich, daß der Wechselwirkungseffekt mit dem Karnofsky-Index wesentlich stärker ausgeprägt ist als der mit dem Alter. Alle Patienten mit Karnofsky-Index > 80 profitieren von der Kombinationstherapie, während bei fast allen Patienten mit Karnofsky-Index < 70 die Monotherapie überlegen ist. Das bedeutet also, daß in diesen beiden Gruppen das Alter für die Therapiewahl praktisch keine Rolle spielt. Lediglich in der mittleren Gruppe, in der sowohl aufgrund dieser Analyse als auch aufgrund der Analyse in Abschnitt 3.1.1. keine eindeutige Therapiepräferenz anzugeben war, scheint für jüngere Patienten die Kombinationstherapie und für ältere Patienten die Monotherapie überlegen zu sein.

Insgesamt ist jedoch noch einmal darauf hinzuweisen, daß dieses Verfahren sehr explorativ ist. Die Ergebnisse bzgl. der „optimalen“ Therapien beruhen allein auf den Punktschätzern der Parameter und sind somit statistisch nicht abgesichert. Sie können jedoch in dem Sinne, die Daten weitergehend zu beschreiben, auch dazu beitragen, die Diskrepanz zwischen den Ergebnissen der Cox- und der CART-Analyse etwas näher zu durchleuchten. Jedoch soll an dieser Stelle nicht näher auf einen Vergleich dieser sehr unterschiedlichen Analyseansätze eingegangen werden, sondern auf eine andere Arbeit, die sich speziell mit diesem Aspekt befaßt, verwiesen werden (SCHMOOR und ULM, 1991).

Zur Präsentation der Analyseergebnisse schlagen EDLER und QUINTERO vor, zusätzlich separat für die beiden Therapiegruppen die aus dem Cox-Modell geschätzten medianen Überlebenszeiten für verschiedene Konstellationen der prognostischen Faktoren einander gegenüberzustellen (ANDERSEN et al., 1983). Dies ist für die durch den Karnofsky-Index definierten prognostischen Untergruppen Abbildung 3.1.2. der Arbeit von ULM et al. zu entnehmen, die die gesamten Überlebenszeitverteilungen darstellt. Dabei wurden die übrigen im Modell enthaltenen Kovariablen nicht auf bestimmte Werte fixiert, sondern innerhalb jeder Untergruppe die geschätzten Überlebensraten gemittelt. Dies ermöglicht, die allein durch den Karnofsky-Index verursachten Unterschiede zu beurteilen, wobei die Schätzungen gleichzeitig bzgl. der anderen prognostischen Faktoren adjustiert wurden.

Literatur

- ANDERSEN, P.K., E. CHRISTENSEN, L. FAUERHOLDT, P. SCHLICHTING (1983): Measuring prognosis using the proportional hazards model. *Scandinavian Journal of Statistics* 10, 49-52.
- ANTES, G. (1990): Multikollinearität in den Einflußvariablen im Coxschen Regressionsmodell für Überlebenszeiten. Dissertation, Universität Bremen.
- SCHMOOR, C., K. ULM (1991): Comparison of the Cox model and the regression tree method in analyzing a controlled clinical trial. Vortrag beim 37. Biometrischen Kolloquium, 19.-22.3.1991 Universität Hamburg.
- SCHUMACHER, M., M. OLSCHESKI, C. SCHMOOR (1987): The impact of heterogeneity on the comparison of survival times. *Statistics in Medicine* 6, 773-784.
- THALL, P.F., J.M. LACHIN (1986): Assessment of stratum-covariate interactions in Cox's proportional hazards regression model. *Statistics in Medicine* 5, 73-83.
- THOMSEN, B.L. (1988): A note on the modelling of continuous covariates in Cox's regression model. Research Report 88/5, Statistical Research Unit, University of Copenhagen.
- TIBSHIRANI, R., T. HASTIE (1987): Local likelihood estimation. *Journal of the American Statistical Association* 82, 559-567.
- ULM, K., C. SCHMOOR, W. SAUERBREI, G. KEMMLER, Ü. AYDEMIR, B. MÜLLER, M. SCHUMACHER (1989): Strategien zur Auswertung einer Therapiestudie mit der Überlebenszeit als Zielkriterium. *Biometrie und Informatik in Medizin und Biologie* 20, 171-205.

Antwort auf den Kommentar von L. Edler und C. Quintero

zu K. Ulm, C. Schmoor, W. Sauerbrei, G. Kemmler, Ü. Aydemir, B. Müller und M. Schumacher: (1989). Strategien zur Auswertung einer Therapiestudie mit der Überlebenszeit als Zielkriterium. Biometrie und Informatik in Medizin und Biologie, 10, 171-205.

W. Sauerbrei

1. Verborgene Kovariable und Variablenselektion

Edler und Quintero sprechen von dem begrenzten medizinischen Wissen über die Krankheit, und damit verbunden ist auch das begrenzte Wissen über prognostische Faktoren. Nach inhaltlichen Gesichtspunkten wurden in der Gliom-Studie mehr als 10 Faktoren diskutiert, deren prognostische Bedeutung umstritten ist. Bei anderen Krankheitsbildern ist diese Zahl häufig noch höher. Würde man bereits bei der Planung einer Studie zumindest gute Kenntnisse über einige starke Faktoren besitzen, so böte sich eine der folgenden Möglichkeiten an:

- i) Unterschiedliche Therapiestudien für Kollektive mit unterschiedlicher Prognose. Dieses „risiko-adaptierte“ Vorgehen würde zu homogenen Populationen führen, bei denen die Effekte weiterer Faktoren nur noch sehr gering sind und somit eine Adjustierung kaum Auswirkungen auf den Therapieeffekt hätte. Dies kann allerdings zu so geringen Patientenzahlen führen, daß eine Therapiestudie in einem vernünftigen Zeitraum kaum durchgeführt werden kann.
- ii) Bereits in dem Studienprotokoll werden die Faktoren festgelegt, bezüglich derer der Therapieeffekt adjustiert wird. Die Faktoren sind dabei möglichst exakt einschließlich des verwendeten Skalenniveaus anzugeben. In entsprechender Weise werden auch prospektiv die zu untersuchenden Interaktionsterme festgelegt. Im konfirmatorischen Sinne können dann die einzelnen Faktoren auf ihre Bedeutung für die Überlebenszeit der Patienten überprüft werden.

Gegenüber der gängigen Praxis wie im Fall der Gliom-Studie hat das zweite Vorgehen den großen Vorteil, daß die häufig kritisierte nachträgliche Variablenselektion entfällt. Für den Fall der Subgruppenanalyse sieht auch Buyse (1989) darin einen sinnvollen und häufig auch praktikablen Ansatz. Die Voraussetzung guter Kenntnisse der Bedeutung prognostischer Faktoren ist sicherlich nicht immer anzutreffen, trotzdem könnte durch eine sorgfältige Planung bei einigen Studien dieses Vorgehen durchgeführt werden. Beispiele sind dafür neuere Studienprotokolle aus dem Bereich der gynäkologischen Onkologie. Die Nachteile der beiden Vorgehen – „nicht optimale“ Adjustierung in dem gegebenen Datensatz bei prospektiver Festlegung des Modells versus retrospektive datenabhängige Modellwahl mit eventuellen Interpretationsproblemen des Therapieeffektes –

sind gegeneinander abzuwägen. In einem Beispiel zeigen Beach und Meier (1989) die Interpretationsprobleme bei einem „grenzwertig signifikanten“ Therapieeffekt auf, wenn dieser stark von der datenabhängigen Modellwahl beeinflusst wird.

Ein wichtiger Aspekt bei der prospektiven Festlegung des Modells ist sicherlich auch die Studiendauer, da zum Zeitpunkt der Auswertung von Langzeitstudien die prognostische Bedeutung eines Faktors bereits etabliert sein kann, der zum Zeitpunkt der Studienplanung noch als „verborgene Kovariable“ zu bezeichnen war.

2. Skalenniveau

Über den ersten wichtigen Schritt der Analyse, die Festlegung des Skalenniveaus der Variablen, wird nur wenig in der Literatur angegeben, und auch Edler und Quintero vermissen eine stärkere Berücksichtigung dieses Aspekts.

Häufig wird die funktionale Form der Abhängigkeit einer Zielgröße von den Einflußgrößen jeweils durch den bivariaten Zusammenhang von Zielgröße und jeweils einer Einflußgröße festgelegt. Oft wird eine Kategorisierung der Einflußgrößen auch durch inhaltliche Aspekte und die zugehörige Verteilungsfunktion ohne die Berücksichtigung der Zielgröße durchgeführt. Um einen Vergleich verschiedener Modellansätze zu ermöglichen, sollte von der Kodierung (Ulm et al, Tab. 2.1) höchstens noch in Form einer „Vergrößerung“ des Skalenniveaus abgewichen werden. Weitere Ansätze aus der Literatur werden in dieser Erwiderung bei SCHMOOR (1991) diskutiert.

Eine andere Möglichkeit bietet der Bootstrap-Ansatz, bei dem die beiden Schritte Kodierung und Selektion prognostischer Faktoren verbunden werden. Da zu Beginn der Analyse aus medizinischer Sicht häufig sehr viele Kategorien betrachtet werden sollen, bietet dieser Ansatz die Möglichkeit, die wirkliche Notwendigkeit der Trennung einzelner Kategorien im Kontext mit den weiteren Variablen zu sehen. Verdeutlicht wird dies in Tabelle 1 am Beispiel der beiden Dummy-Variablen für den Malignitätsgrad.

Die gemeinsame Verteilung zeigt, daß der Malignitätsgrad in allen Bootstrap-Replikationen als prognostischer Faktor erkannt wurde. Die Variable Dummy 1 scheint eine viel stärkere prognostische Bedeutung zu haben als die Variable Dummy 2. Dies stimmt mit den Kaplan-Meier-Schätzern (ULM et al., Abb. 2.2) gut überein. Die zusätzliche prognostische Bedeutung der Variablen Dummy 2 kann eher als sehr fraglich angesehen werden, da sie nur in weniger als einem Drittel der Modelle zusätzlich zur Variablen

Tabelle 1: Gemeinsame Verteilung der Aufnahmehäufigkeit der beiden Dummy-Variablen des Malignitätsgrades.

		Dummy 2 (3 und 3-4 vs 4) im Modell		
		ja	nein	
Dummy 1 (3 vs 3-4 und 4) im Modell	ja	26 (29,5)	62 (70,5)	88
	nein	12 (100,0)	0 (0)	12
		38	62	100

Dummy 1 aufgenommen wurde. Wir sind uns bewußt, daß natürlich die durch die Definition der beiden Dummyvariablen erzeugte Korrelation fälschlicherweise eine Aufnahme der zweiten Variable ins Modell verhindern kann. Dies wurde bewußt in Kauf genommen, da im Gegensatz zu der Analyse in Abschnitt 3.1. stärker das Ziel verfolgt wurde, ein einfacheres Modell zur Adjustierung des Therapieeffektes zu entwickeln.

3. Variablen Cluster

Das Beispiel von Edler und Quintero zur Bestimmung prognostischer Faktoren beim fortgeschrittenen kolo-rektalen Karzinom kann man auch als ein Variablen Cluster Problem (HARRELL et al. 1984) ansehen. Untersucht wird die Frage, ob „weiße Blutkörperchen“, gemessen in der Form Leukozyten und Granulozyten, eine prognostische Bedeutung haben. Enthält das Modell beide hochkorrelierte Variablen, so erreicht keiner der beiden standardisierten Koeffizientenschätzer den kritischen Wert 1.96 für $\alpha = 5\%$. Bei einem schrittweisen Vorgehen werden die Leukozyten als Vertreter der Variablen „weiße Blutkörperchen“ ausgewählt. Das Beispiel verdeutlicht zusätzlich, daß Schätzer und Varianz der weiteren Variablen nur gering von den gewählten Modellparametern des Variablen Clusters abhängen.

Eine weitere Möglichkeit der Einbeziehung von Variablen Clustern besteht in der Bildung einer neuen Variablen, z.B. einem geeigneten Index, die das Cluster repräsentieren soll. Als ein in der Onkologie bekanntes Beispiel für dieses Vorgehen kann das histologische Grading nach Bloom & Richardson angesehen werden, bei dem aus drei Dimensionen ein Score zwischen 3 und 9 Punkten gebildet wird, der dann nochmals zu der Dreier-Gruppierung des Grading zusammengefaßt wird.

4. Signifikanzniveau

Edler und Quintero sprechen die Wahl des Signifikanzniveaus α^* für die Aufnahme von weiteren Variablen an. Ihr Vorschlag, α^* in der univariaten Analyse kleiner zu wählen, führt dazu, insgesamt weniger Faktoren zu betrachten, aber es löst nicht das Problem, mit korrelierten Faktoren zu starten. Die Folge ihres Vorschlages kann sein, daß z.B. von zwei hochkorrelierten Faktoren mit großer prognostischer Bedeutung beide für den nächsten Schritt im Modell verbleiben, während unkorrelierte Faktoren mit schwacher prognostischer Bedeutung bereits an der „ersten Hürde“ des sehr niedrigen α^* -Levels scheitern.

Für die Eliminierung korrelierter Faktoren schlagen wir stattdessen Subset-Selektions-Methoden vor. Schrittweise (Forward, Backward, Stepwise) Selektion oder All-Subsets-Ansätze mit unterschiedlichen Kriterien (z.B. Akaike's Informationskriterium) sind sicherlich ein Instrument, um das Problem korrelierter Faktoren zu behandeln. Wir sind uns vieler Kritikpunkte dieser Ansätze bewußt. Bei den schrittweisen Prozeduren ist z.B. die Frage geeigneter kritischer Werte für die weitere Aufnahme bzw. Elimination von Variablen offen. Diese Werte werden vorgegeben. Aussagen zu dem Signifikanzniveau der im Endmodell berücksichtigten Faktoren sind nicht möglich. Für weitere Aspekte möchten wir auf den guten Übersichtsartikel von MILLER (1984) verweisen.

Auch bei dem eigenen Bootstrap-Ansatz kann mit Hilfe des Trennungskriteriums für die Aufnahmehäufigkeit (bei ULM et al. 30%) eine Steuerung eingeführt werden, um eher „viele“ oder „wenige“ Variablen für den nächsten Schritt aufzunehmen. Aussagen zum Signifikanzniveau sind aber auch hier nicht möglich.

5. Analyse von Prognosestudien

In diesem Abschnitt möchten wir die Aussage von Edler und Quintero relativieren, daß „die vorgestellten Methoden erheblichen Modellcharakter für die Auswertung von Prognosestudien“ haben. Auch wir sind der Überzeugung, daß die Qualität der Daten – ein immer noch stark vernachlässigtes Problem in der Medizin – in Therapiestudien besser ist als bei retrospektiven Fallsammlungen.

Während zur Adjustierung des Therapieeffektes auch schwache prognostische Faktoren ins Modell aufgenommen werden sollen, ist dieses Vorgehen bei Prognosestudien nicht unbedingt angebracht. Dies unterschiedliche Ziel hat auch Auswirkungen auf die Analysestrategie.

Bei Prognosestudien können wir uns vor allem die folgenden typischen Ausgangssituationen vorstellen:

- i) Interesse an wirklich starken Faktoren, evtl. auch Abklärung, welcher von zwei oder mehreren möglicherweise korrelierten Faktoren der wichtigere ist. Die Faktoren können zum Beispiel als Grundlage für die Studienpopulationen von risikoadaptierten Therapiestudien dienen.
- ii) Studien, die vor allem an der Prognose einzelner Individuen interessiert sind. Auch bei Diagnosestudien ist ein vergleichbares Ziel vorhanden. In diesem Fall versucht man eine möglichst gute Anpassung bezüglich anderer Kriterien, z.B. den „Mean Square Error (of Prediction)“ oder die „Trefferquote“, zu erreichen. Eine Übersicht zu geeigneten Zielkriterien bei Modellen zur Vorhersage von Überlebenszeiten findet sich bei KORN und SIMON (1990).

6. Bootstrap-Resampling

Edler und Quintero sprechen die Möglichkeit der Modelldiagnostik mit Hilfe des Bootstrap-Ansatzes an. Auch wir sehen darin ein weiteres sinnvolles Anwendungsspektrum der Bootstrap-Idee.

Wie bereits bei ULM et al. bemerkt, sind in jeder Bootstrap-Replikation im Mittel über 35 % der Patienten nicht enthalten.

Sei $x^* = (y^*, z_1^*, \dots, z_k^*)$ ein einflußreicher Punkt für die j -te Komponente des Parametervektors. Diese Beobachtungseinheit ist in mehr als einem Drittel der Replikationen nicht enthalten, in anderen teilweise sogar mehrfach, d.h. ihr Einfluß wird sich noch wesentlich verstärken. Die aus den Bootstrap-Replikationen geschätzte Verteilung des Parameters β_j wird im wesentlichen zweigipflig sein: Nahezu eine Normalverteilung für die Replikationen, in denen der einflußreiche Punkt x^* nicht enthalten ist und mit einem großen Unterschied in der Lage und einer schiefen Verteilung für die Replikationen, in denen x^* mindestens einmal enthalten ist.

Insofern können die Bootstrap-Verteilungen der geschätzten Parameter einen wichtigen Hinweis auf einflußreiche Punkte liefern. Existiert keine oben beschriebene zweigipflige Verteilung, so kann dies als eine Überprüfung auf einflußreiche Punkte für einzelne Komponenten des Parametervektors angesehen werden.

Erkennt man aber eine solche Form, so wird die Identifikation des einzelnen Punktes relativ kompliziert und sehr rechenzeitaufwendig sein. Zunächst müssen die Bootstrap-Replikationen mit Extremwerten der β_j identifiziert werden. Dies gelingt sicherlich leicht. Für diese Replikationen müssen aber Gemeinsamkeiten bezüglich der Aufnahmehäufigkeit der Beobachtungseinheiten untersucht werden. Hilfreich können dabei (Rang)-Korrelationsuntersuchungen von $\hat{\beta}_j$ und der Aufnahmehäufigkeit der Patienten in den Bootstrap-Replikationen sein. Dies ist für jeden Patienten durchzuführen.

Diese Überlegungen gelten zunächst für die Schätzung des vollen Modells. Bei der Variablenselektion kann der einflußreiche Punkt x^* dazu führen, daß die Variable z_j nur in den Replikationen ins Modell aufgenommen wird, in denen x^* auch (mehrfach) in der Bootstrap-Replikation enthalten ist. Variablen, deren Aufnahmehäufigkeit über 70% liegt, können somit nicht nur durch einen einflußreichen Punkt so bedeutend erscheinen.

Im Beispiel der Gliom-Daten wurden diese Ansätze lediglich zur Überprüfung des Endmodells eingesetzt. Keine der auf 500 Replikationen basierenden Verteilungen der Parameterschätzer wich wesentlich von einer Normalverteilung ab. Wir schließen daraus, daß für dieses Modell keine einflußreichen Punkte im Datensatz enthalten sind.

Diese heuristischen Überlegungen bilden ein interessantes weiteres Gebiet der Modelldiagnostik, ihr konkreter Einsatz erfordert aber neben weiterer theoretischer Abklärung auch in viel stärkerem Maße als alle bisherigen Vorschläge Überlegungen zur „optimalen“ Programmierung. Auch Rechenzeitaspekte können dabei nicht vollkommen ignoriert werden.

Edler und Quintero weisen auf eine Reihe weiterer Möglichkeiten zur Modelldiagnostik hin, die aber in der Praxis vermutlich immer noch zu selten eingesetzt werden. Ein Grund dafür ist sicherlich im (teilweisen) Fehlen dieser Ansätze in den Statistikprogrammpaketen zu sehen.

Auch wir sehen in der von Edler und Quintero vorgeschlagenen Untersuchung der Korrelationsmatrix der Kovariablen eine sinnvolle Möglichkeit der Untersuchung der Kollinearität.

Ein vollkommen anderer Weg wird durch die Untersuchung der 2x2-Tafeln der Aufnahmehäufigkeit der Variablen im Bootstrap-Ansatz gewählt. Dieser Vorschlag berücksichtigt den Einfluß der Kovariablen auf die Zielgröße und ist somit als eine zusätzliche Möglichkeit der Untersuchung der Kollinearität anzusehen.

Der sonst übliche Einsatz von Bootstrap-Verfahren bezieht sich vor allem auf Schätzungen zur Streuung von Parametern. Für die Bestimmung von Konfidenzintervallen sind Replikationen von 1000 bis 2000 sicherlich nötig. ULM et al. haben Konfidenzintervalle für den Therapieeffekt im Endmodell aufgrund von 500 Replikationen geschätzt. Diese Zahl bildet vermutlich die untere Grenze für die Bestimmung von Konfidenzintervallen, zumindest wenn die Perzentilmethode verwendet wird. Die Schätzung des Konfidenzintervalls stand aber nicht im Mittelpunkt der Arbeit, sie sollte lediglich den Bootstrap-Ansatz abrunden. Weitere Angaben zu Bootstrap-Konfidenzintervallen finden sich in der bei ULM et al. zitierten Literatur.

Eine andere Situation liegt im Fall der Aufnahmehäufigkeit von Variablen vor. Im ersten Schritt des Bootstrap-Selektionsverfahrens sind wir nur an einem Punktschätzer der Binomialverteilung der Aufnahmehäufigkeit interessiert.

Die Stabilität der Aufnahmehäufigkeiten in Abhängigkeit von der Anzahl der Bootstrap-Replikationen ist in Tabelle 2 dargestellt. Eine gute Übereinstimmung für die einzelnen Bootstrap-Analysen ist erkennbar.

Da bei der Analyse der Gliom-Daten auch schwache Faktoren ins Modell aufgenommen werden sollten, wurde im ersten Schritt der Bootstrap-Selektionsprozedur der niedrige Grenzwert 30% gewählt. Bis auf einige Abweichungen bei schwachen Einflußgrößen erhalten wir bei den 6 Bootstrap-Analysen das in der Arbeit von ULM et al. postulierte Modell mit 7 Faktoren.

Die Stabilität in Abhängigkeit von der Anzahl der Bootstrap-Replikationen nimmt noch weiter zu, wenn wir nur starke Faktoren auswählen wollen. Bei einer Veränderung der Grenzwerte im ersten Schritt auf 60% erhalten wir, mit einer Ausnahme bei nur 50 Bootstrap-Replikationen, in allen Fällen das Modell mit den 4 Faktoren Alter, Resektion (Dummy 1), Malignität (Dummy 1) und Karnofsky-Index (Dummy 1).

Tabelle 2: Stabilität der relativen Aufnahmehäufigkeiten (Backward Elimination Methode, $\alpha = 5\%$) in Abhängigkeit von der Anzahl der Bootstrap-Replikationen. Endmodelle bei 2 verschiedenen Cutpoints. (*: Verschiedene Vorzeichen der geschätzten Koeffizienten in den selektierten Modellen. In allen Fällen dominiert ein Vorzeichen.)

		Anzahl der Bootstrap-Replikationen					
Variable		50	100	100	200	400	1000
Alter	Z_5	100	100	100	99.5	100.0	99.9
Resektion							
Dummy 1: partielle Res.	Z_8	100	100	99	100.0	99.7	99.6
Dummy 2: totale Res.	Z_9	36	35*	32	28.0	35.5*	34.9
Malignitätsgrad							
Dummy 1: 3,3-4 und 4	Z_3	88	88	86	89.0	87.5	86.6
Dummy 2: 3 und 3-4,4	Z_4	46	38	38	34.0*	35.2*	37.2*
Karnofsky-Index							
Dummy 1: $< 70, \geq 70$	Z_6	76	66	65	74.0	68.7	73.1
Dummy 2: $\leq 80, > 80$	Z_7	6*	13*	10	7.0*	9.5*	8.0*
Organisches Psychosyndrom	Z_{14}	48*	56	45	41.0	39.0	43.7*
Epilept. Anfälle	Z_{12}	38	52	54	39.5	45.0	41.5
Cortison	Z_{11}	64	49	57	51.5	58.2	56.5
Geschlecht	Z_1	18	33	32	24.5	27.2*	22.8*
Aphasie	Z_{15}	12	23*	25*	15.0	16.7*	17.4*
Anamnesedauer	Z_2	12*	19	19	15.0	18.0*	17.5*
Anti-Epileptika	Z_{10}	12*	19*	14*	19.5*	15.7	14.7*
Bewußtseinsstörung	Z_{13}	14*	17*	16	13.0*	19.0*	16.4*
Modell mit Strategie A und Cutpoint 0.30		(M) + Z_4	(M)	(M)	(M) - Z_{12} - Z_{14}	(M) + Z_9	(M) + Z_9
(M) = $Z_3 + Z_5 + Z_6 + Z_8 + Z_{11} + Z_{12} + Z_{14}$							
Modell mit Strategie A und Cutpoint 0.60		(N) + Z_{11}	(N)	(N)	(N)	(N)	(N)
(N) = $Z_3 + Z_5 + Z_6 + Z_8$							

Trotz unserer geringen Erfahrung glauben wir, daß bereits ab 100 Replikationen das Bootstrap Resampling-Verfahren sinnvoll eingesetzt werden kann. Untersuchungen zur Stabilität von Modellen werden vermutlich viel zu selten durchgeführt, zumindest werden sie nicht häufig in der biometrischen Literatur beschrieben. ULM et al. haben durch den Vergleich unterschiedlicher Analysestrategien einen Versuch unternommen, für den konkreten Fall einer Therapiestudie die Abhängigkeit der inhaltlichen Aussagen von dem methodischen Vorgehen zu untersuchen.

Literatur

- BEACH, M.L., MEIER, P. (1989): Choosing Covariates in the Analysis of Clinical Trials, *Controlled Clinical Trials*, 10, 161-175
- BUYSE, M.E. (1989): Analysis of Clinical Trial Outcomes: Some Comments on Subgroup Analysis, *Controlled Clinical Trials*, 10, 187-194
- HARREL, F.E., LEE, K.L., CALIFF, R.M. et al. (1984): Regression modelling strategies for improved prognostic prediction, *Statistics in Medicine*, 3, 143-152
- KORN, E.L., SIMON, R. (1990): Measures of Explained Variation for Survival Data, *Statistics in Medicine*, 9, 487-503
- MILLER, A.J. (1984): Selection of Subsets of Regression Variables, *Journal of the Royal Statistical Society*, 147 (3), 389-425
- SCHMOOR, C. (1991): Antwort auf den Kommentar von L. Edler und C. Quintero, *Biometrie und Informatik in Medizin und Biologie* 22, 69-76.

Anschrift des Verfassers: Dipl.-Stat. Willi Sauerbrei, Institut für Medizinische Biometrie und Informatik, Universität Freiburg, Stefan-Meier-Straße 26, 7800 Freiburg i. Br.

Antwort auf den Kommentar von L. Edler und C. Quintero

zu K. Ulm, C. Schmoor, W. Sauerbrei, G. Kemmler, Ü. Aydemir, B. Müller und M. Schumacher: (1989). Strategien zur Auswertung einer Therapiestudie mit der Überlebenszeit als Zielkriterium. Biometrie und Informatik in Medizin und Biologie, 10, 171-205.

W. Vach

1. Behandlung fehlender Werte

Die Behandlung fehlender Werte in einer Kovariablen als eigene, zusätzliche Kategorie ist naheliegend und im Rahmen der Auswertung epidemiologischer Studien auch verbreitet. Die Einführung eines zusätzlichen Parameters (für jede betroffene Kovariable) erscheint dabei im Vergleich zur sonst notwendigen Reduzierung der Anzahl der auswertbaren Fälle oder der berücksichtigbaren Variablen vertretbar. Allerdings zeigt eine genauere Betrachtung, daß dieser Ansatz im Coxschen Regressionsmodell nicht verträglich mit den Modellannahmen ist. Dies ist selbst dann der Fall, wenn die Werte zufällig fehlen, d.h. wenn die bedingte Wahrscheinlichkeit für das Auftreten eines fehlenden Wertes in einer Kovariablen weder von der wahren Ausprägung dieser Kovariablen, noch von der der anderen Kovariablen, noch von der wahren oder zensierten Überlebenszeit abhängt. In diesem Fall können wir die Subpopulation mit fehlenden Werten in einer Kovariablen als zufällige Stichprobe aus dem Gesamtkollektiv ansehen. Für diese Subpopulation mit unbekannter Ausprägung der Kovariablen ergibt sich die Hazardfunktion als Mischung der Hazardfunktionen für die verschiedenen Kategorien der Kovariablen, wobei die Mischungsparameter zeitabhängig sind (SCHUMACHER, OLSCHESKI und SCHMOOR 1987).

Damit ist jedoch das relative Risiko zwischen der neu gebildeten und den übrigen Kategorien der Kovariablen als auch zwischen Kategorien anderer Kovariablen innerhalb der Subpopulation nicht konstant über die Zeit, also eine wesentliche Voraussetzung des Coxschen Regressionsmodells verletzt. Es ist daher nicht auszuschließen, daß wir verzerrte Schätzer erhalten. Dabei ist zu erwarten, daß die Verzerrung mit dem Effekt der betroffenen Kovariablen wächst und die Gleichverteilung der einzelnen Kategorien die ungünstigste Verteilung ist, da hier die Abweichung vom konstanten relativen Risiko am größten ist.

Über die Größenordnung der möglichen Verzerrung ist bisher nichts bekannt. Erste Anhaltspunkte ergeben sich aus vergleichbaren Untersuchungen für das logistische Regressionsmodell (VACH & BLETTNER, in Vorbereitung). Dort wird durch die Behandlung fehlender Werte als eigene Kategorie die Annahme eines konstanten Odds Ratios ebenfalls verletzt, und es können erhebliche Verzerrungen auftreten.

Für eine erste Untersuchung der möglichen Auswirkungen des Ansatzes im Cox'schen Regressionsmodell haben wir für die 413 Patienten, die den Schätzungen von Tabelle 3.1.4 in ULM et al. (1989) zugrunde liegen, künstlich fehlende Werte in verschiedenen Variablen erzeugt und Schätzungen der Effekte unter Einführung einer zusätzlichen Kategorie durchgeführt. Zuerst wurden für die Variablen *Resektion*, *Malignitätsgrad*, *Epileptische Anfälle* und *Cortison* jeweils einzeln für ein Viertel der Patienten die Ausprägung auf „missing“ gesetzt. Die neuen Schätzungen zeigten keine auffälligen Abweichungen von denen aus Tabelle 3.1.4; die Abweichungen bewegten sich alle in einer Größenordnung wie sie auch bei einer zum Vergleich durchgeführten Schätzung basierend auf den vollständigen Fällen beobachtet werden konnten. Auch wenn in allen vier Variablen gleichzeitig fehlende Werte erzeugt wurden, veränderten sich die Schätzer kaum. Die Schätzer für die Koeffizienten der neu gebildeten Kategorien waren dabei erwartungsgemäß betragsmäßig kleiner als jene für den Unterschied zwischen den ursprünglichen Kategorien.

Da wir keine auffälligen Verzerrungen beobachten konnten, schien es uns gerechtfertigt, mit diesem Ansatz auch eine Berücksichtigung der Variable Tumorrest zu versuchen. Nach Aufnahme der Variable in das Modell ergab sich bei einer Schätzung basierend auf den verbleibenden 252 vollständigen Fällen für den Tumorrest ein geschätzter Koeffizient von 0.530 (S.E. 0.223) und ein p-wert von 0.015.

Wurden nun fehlende Werte in der Variablen Tumorrest als eigene Kategorie behandelt, so standen wieder alle 413 Patienten zur Verfügung. Für den Tumorrest ergab sich nun ein Koeffizient von 0.465 (S.E. 0.207) und für die neue Kategorie 0.410 (S.E. 0.206). Der p-wert ergab sich zu 0.055. Hier erhebt sich nun der Verdacht, daß die Modellverletzung beruhend auf der Behandlung fehlender Werte als eigene Kategorie zu einer Unterschätzung des Risikos führte, die eine Nicht-Signifikanz bei der Betrachtung der 413 Patienten bewirkte, während die unverzerrte Schätzung aufgrund von 252 Patienten eine positive Signifikanzaussage ermöglichte. Die Ergebnisse hinsichtlich der anderen Variablen wichen kaum von denen in Tab. 3.4.1 ab.

Der Aussagewert einer derartigen, exemplarischen Untersuchung ist natürlich beschränkt. Notwendig ist die genauere Untersuchung der Verzerrungen, die durch die Verletzung der Modellannahmen entstehen kann. Vor dem Abschluß derartiger Untersuchungen kann aber bei dem derzeitigen Kenntnisstand, insbesondere angesichts der für das logistische Regressionsmodell bekannten Ergebnisse, keine Empfehlung für die Behandlung fehlender Werte als eigene Kategorie ausgesprochen werden.

Literatur

- SCHUMACHER M., OLSCHESKI, M. und SCHMOOR, C. (1987): The Impact of Heterogeneity on the Comparison of Survival Times. *Statistics in Medicine* 6, 773-784
- VACH, W. und BLETNER, M. (1991): Biased Estimation of the Relative Risk in Case-Control Studies due to the use of Ad-Hoc Methods for Missing Values in Confounding Variables. *American Journal of Epidemiology* (im Druck).

Anschrift des Verfassers: Dipl.-Stat. Werner Vach, Institut für Medizinische Biometrie und Informatik, Universität Freiburg, Stefan Meier Str. 26, 7800 Freiburg

BUCHBESPRECHUNG/BOOKREVIEWS

GÖTTSCHE, T.

Einführung in das SAS-System für den PC

Gustav Fischer Verlag, Stuttgart 1990

ISBN 3-437-40233-1. 298 S. DM 49,-

Das vorliegende Buch ist ausgerichtet auf einen Leserkreis ohne Vorkenntnisse im Umgang mit dem SAS-System, unabhängig davon, ob der Leser bereits mit anderen Statistikpaketen vertraut ist. Allerdings sollten Grundkenntnisse in der Statistik vorhanden sein.

Die bildliche Darstellung der SAS-Benutzeroberfläche und kurze präzise Erläuterungen ermöglichen es dem Leser, im Selbststudium, innerhalb kurzer Zeit erste SAS-Jobs abzuschicken. Dabei wird der Benutzer von Beginn an dazu angehalten, durch Belegung von Funktions- und Buchstabentasten die Befehlseingabe abzukürzen.

Im Anschluß hieran geht der Autor auf die Möglichkeit des Datenmanagement mit SAS ein. Die Leitlinie von SAS, die Bearbeitung der Daten durch eigenständige Programmierung und modulare Nutzung der SAS-Statistikprozeduren wird dabei vorzüglich vermittelt. Dieser Abschnitt mag auch für Leser von Interesse sein, die bereits einen Einstieg in das SAS-System Gewonnen haben. Demgegenüber hätten die Ausgabemöglichkeiten unter SAS besser dargestellt werden können.

Die Behandlung der Statistikprozeduren beschränkt sich neben der deskriptiven Statistik auf Korrelation und univariate Varianzanalyse sowie gebräuchliche nichtparametrische Verfahren. Die Einführung in die Nutzung der Prozeduren erfolgt anhand von Beispielen aus Statistikbüchern. Für weitergehende Fragestellungen und spezifische Probleme ist hier die Schnittstelle zum SAS/STAT User's Guide zu sehen. An die aufgeführten Statistikprozeduren knüpft die Darstellung der Graphikprozeduren an, womit der Anwender in der Lage ist, seine zuvor gewonnenen Ergebnisse in einer Präsentationsgraphik darzustellen. Für ausgefeilte Graphiken mag der Anwender auf den SAS/GRAPH User's Guide zurückgreifen.

Das Buch wird damit der Zielsetzung, dem Leser das grundlegende „Handwerkszeug“ für den Umgang mit dem SAS-System zu liefern, voll gerecht.

W. Grüneberg, Hannover

MERKEL, B., SPERLING, B.

Statistik für Micro-Computer

mit dem Statistik-Programm STASY 100

Benutzerhandbuch und Programmdiskette für IBM PC/XT/AT & Kompatible unter DOS

Gustav Fischer Verlag, Stuttgart 1990

ISBN 3-437-40231-5, 151 S. mit Diskette (1,2 Mb; 5 1/4"), kt. DM 48,-

Wie Broder Merkel und Barbara Sperling in Ihrem Vorwort schreiben, wollen sie mit ihrem Buch (in Verbindung mit dem zugehörigen STASY-100 Programm der PIC GmbH für PC unter DOS) vor allem solche Leute, die „Vorurteile und Berührungsängste“ vor bzw. bei der Anwendung von Statistik haben, ansprechen. Ihr Buch erhebt nicht den Anspruch, ein Lehrbuch der Statistik zu sein, sondern es soll „Anfängern helfen, sich in die Grundbegriffe der Statistik unter Benutzung der EDV einzuarbeiten“. Es soll auch verdeutlichen, daß die Anwendung von Statistik „durchaus Spaß machen kann“.

Diesem Anspruch wird das Buch durchaus gerecht. Es basiert auf einer Art „Learning by Doing“. Wenn man es (am PC sitzend!) von Anfang an durcharbeitet und dabei die gestellten Übungsaufgaben bzw. die jeweils direkt darauf folgenden Lösungen nachvollzieht (viel mehr ist kaum möglich, da die detaillierte Programmbeschreibung erst später im 2. Teil des Buches nach dem Statistikteil erfolgt), erhält man sehr schnell „faßbare“ Ergebnisse, wenn z. B. die auf der Diskette mitgelieferten Beispieldatensätze graphisch dargestellt und durchleuchtet werden. So werden einem – quasi spielerische und wirklich motivierende Art und Weise – allmählich sowohl die Funktionsweise des Programmes STASY-100 wie auch das Prinzip statistischer Analysen (wenn es so etwas überhaupt gibt) vertraut.

In der Einleitung erfolgt zunächst eine kurze Einführung in das Arbeiten mit dem PC, wo z. B. auch grundlegende DOS-Befehle kurz beschrieben werden. Es folgen eine Einführung in das interaktive Statistik-Programm STASY-100 und Anweisungen für die Installation. Ansonsten teilt sich das Buch in zwei Hauptteile, das mehr theoretische Statistik-Tutorial im Teil A und die Programmbeschreibung im Teil B des Buches. Dabei deckt sich die Kapitel-Numerierung des statistischen Teils des Buches weitgehend mit der Menue-Numerierung des STASY-100.

Obwohl die Autoren mit ihrem Statistik-Programm STASY-100 nicht den Anspruch erheben, (Standard-) Statistik-Pakete ersetzen zu wollen, ist es schade, daß mit STASY-100 die Möglichkeiten Daten zu transformieren (z. B. Body-Mass-Index aus Größe und Gewicht zu berechnen) oder das Selektieren von Untergruppen (z. B. von Geschlechts- und Altersgruppen) *sehr* beschränkt sind. Zur Numerik des Programms: Ich habe aus Interesse einige STASY-100-Analysen parallel auch mit SAS berechnet. Die Ergebnisse waren identisch.

Das Buch ist (worauf die beiden Autoren wie gesagt selber in der Einleitung hinweisen) weder dafür geeignet, eine fundierte Einführung in die Statistik zu ersetzen (dafür ist es stellenweise zu wenig exakt), noch dafür, als eine geeignete alleinige Einführung in das Arbeiten mit dem PC zu dienen. Als *zusätzliches* motivierendes (!) Hilfsmittel beim Erlernen von (praktischer) Statistik unter Benutzung eines PC ist das Buch empfehlenswert. Mir hat es Spaß gemacht, es durchzuarbeiten.

Kira Schulz, Neuherberg

O. RICHTER, D. SÖNDGERATH

Parameter Estimation in Ecology · The Link between Data and Models

VCH Verlagsgesellschaft Weinheim 1990

ISBN 3-527-27954-7, IX 218 S., 59 Abb., 35 Tab., DM 158,-

Die Autoren (Prof. Dr. Otto Richter und Dr. Dagmar Söndgerath) arbeiten am Institut für Geographie und Geoökologie der Technischen Universität Braunschweig. Der im Untertitel angestrebte Brückenschlag zwischen Daten und Modellen ist ihnen hervorragend gelungen. Das in seiner Themenauswahl originelle Buch stellt eine Pionierleistung dar. Die fünf Kapitel tragen folgende Überschriften: Einleitung, Mathematische und statistische Hilfsmittel, Modelle in der Form von Wahrscheinlichkeitsverteilungen, Diskrete Bevölkerungsmodelle, Stetige Ökosystemmodelle. Außerdem gibt es drei Anhänge über Wahrscheinlichkeitsrechnung und Statistik, Konvergenz iterativer Verfahren und Berechnung von Maximum-Likelihood-Schätzern für die negative Binomialverteilung.

Das Buch wendet sich an Wissenschaftler, die im Gebiet der biologischen Modellierung, der Biometrie, der Biomathematik und der Angewandten Statistik arbeiten. Von den biologischen Lesern werden mathematische Vorkenntnisse erwartet, insbesondere Grundlagen der Statistik und Infinitesimalrechnung. Das Buch enthält viele Anwendungsbeispiele, vorwiegend aus dem landwirtschaftlichen Bereich. Für die Illustration der Methoden werden nur reale Daten verwendet. Das Buch basiert auf einem einsemestrigen Kurs über statistische Methoden in der Ökologie, der an der Universität Bonn abgehalten wurde.

Im Kapitel über mathematische und statistische Hilfsmittel ist besonders der Abschnitt über Differentialgleichungsmodelle nützlich. An einem Lotka-Volterra-Modell wird der Einsatz des Programmpaketes BMDP zur Parameterschätzung erläutert. Außerdem gibt es Hinweise auf das Programm PARFIT von H. G. Bock.

Das dritte Kapitel ist der statistischen Behandlung von Punktprozessen in der Ebene und der Analyse von Überlebenszeitdaten gewidmet. Kapitel 4 gibt eine kurze Einführung in die Theorie der Leslie-Prozesse und Kapitel 5 behandelt Wachstumsmodelle für Pflanzen und das Lotka-Volterra-Modell.

Das Buch füllt eine Lücke im methodischen Schrifttum. Es ist leider zu befürchten, daß es wegen seines hohen Preises keine weite Verbreitung finden wird.

K. Dietz, Tübingen

TABLE OF CONTENTS

Utilisation de données intermédiaires pour corriger la prédiction de modèles mécanistes	R. Faivre, B. Goffinet, et D. Wallach	1
Increasing precision or reducing expense in regression experiments by using information from a concomitant variable	B. Engel and P. Walstra	13
Extension of Dunnett's multiple comparison procedure to the case of several groups	S. H. Cheung and B. Holland	21
Generalized variance component models for clustered categorical response variables	M. E. Miller and J. R. Landis	33
Multiple comparisons for analyzing dichotomous response	W. W. Piegorsch	45
On estimating HLA/disease association with application to a study of aplastic anemia	S. G. Self, G. Longton, K. J. Kopecky, and K.-Y. Liang	53
Selecting an efficient design for assessing exposure-disease relationships in an assembled cohort	S. Wacholder, M. Gail, and D. Pee	63
On dependent repeated screening tests	T. Lau	77
Sensitivity analysis for matched case-control studies	P. R. Rosenbaum	87
The problem of a covariate-time qualitative interaction in a survival study	J. O'Quigley and F. Pessione	101
Nonparametric tests of association between survival time and continuously measured covariates: The logit-rank and associated procedures	J. O'Quigley and R. L. Prentice	117
Running average analysis of clinical trial ambulatory blood pressure data	M. R. Crager and M. A. Reitman	129
Characterization of dose-response relationships inferred by statistically significant trend tests	R. L. Kodell and J. J. Chen	139
Tests and model selection for the general growth curve model	J. C. Lee	147
Unequally spaced longitudinal data with AR(1) serial correlation	R. H. Jones and F. Boadi-Boateng	161
Estimation of spermarche from longitudinal spermaturia data	M. Jørgensen, N. Keiding, and N. E. Skakkebaek	177
Reduction of selection differentials in finite populations with a nested full-half sib family structure	T. H. E. Meuwissen	195
Multistage index selection in finite populations	L. Norell, T. Arnason, and K. Hugason	205
A random field approach to the analysis of field-plot experiments and other spatial experiments	D. L. Zimmerman and D. A. Harville	223
A bivariate distribution approach to modeling forest diameter distributions at two points in time	B. R. Knoebel and H. E. Burkhart	241
Quantifying precision of mark-recapture estimates using the bootstrap and related methods	S. T. Buckland and P. H. Garthwaite	255
A nonparametric approach to size-biased line transect sampling	P. X. Quang	269
Shorter Communications		
A nonproportional hazards Weibull accelerated failure time regression model	K. M. Anderson	281
Repeated confidence intervals for a scale change in a sequential survival study	D. Y. Lin and L. J. Wei	289
Estimating relative potency using prior information	P. T. Kim, E. M. Carter, and J. J. Hubert	295
Robustness and balance in the mixed model	D. M. Roche	303
Detecting cluster in a heterogeneous community sampled by quadrats	A. R. Solow and W. Smith	311
Estimating standardized risk differences from odds ratios	S. Greenland and P. Holland	319
The Consultant's Forum		
Reader Reaction: Predictive probability in clinical trials	A. P. Grieve	323
Response	S. C. Choi and P. A. Pepple	330
Ethical aspects of statistical practice	D. J. Finney	331
Correspondence		341
Book Reviews		345
Corrections		357
Papers to be published		359